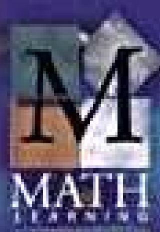
Richard L. Burden
J. Douglas Faires

Análisis numérico



Séptima edición

Contenido

4 Diferenciación e integración numéricas 166

5 Problemas de valor inicial para ecuaciones diferenciales ordinarias 249

6 Métodos directos para resolver sistemas lineales 344

7 Métodos iterativos en el álgebra matricial 417

8 Teoría de la aproximación 483

9 Aproximación de los valores característicos 550

10 Soluciones numéricas de sistemas de ecuaciones no lineales 600

11 Problemas con valor en la frontera para ecuaciones diferenciales ordinarias 645

12 Soluciones numéricas para las ecuaciones diferenciales parciales 691

Prefacio

Acerca del texto

Hemos elaborado este material para una serie de cursos acerca de la teoría y aplicación de las técnicas de aproximación numérica. Está diseñado sobre todo para estudiantes orientados a las matemáticas, ciencias e ingeniería que han concluido su curso de cálculo en licenciatura. Será de utilidad estar familiarizado con los fundamentos del álgebra de matrices y las ecuaciones diferenciales, aunque en el texto presentamos un material introductorio adecuado para estos temas, de modo que estos cursos no son prerrequisitos.

Las ediciones anteriores de *Análisis numérico* se utilizan en situaciones muy variadas. En algunos casos, se enfatizó el análisis matemático en que se basa el desarrollo de las técnicas de aproximación y no los propios métodos; en otros, el énfasis fue a la inversa. Asimismo, el libro se utiliza como referencia básica para cursos iniciales posteriores a la licenciatura en programas de ingeniería y ciencias de la computación; como base para un examen de actuaría en métodos numéricos, donde es común el estudio autodidacta; y en cursos de análisis introductorio impartidos en universidades internacionales. Hemos tratado de adaptar el libro a estos usuarios tan diversos sin comprometer nuestro propósito original:

> *Ofrecer una introducción a las técnicas modernas de aproximación; explicar cómo, por qué y cuándo se espera que funcionen; y proporcionar una base firme para el estudio posterior del análisis numérico y el cómputo científico.*

El libro contiene suficiente material para un año completo de estudio, aunque tal vez los lectores lo utilicen sólo para un curso de un semestre. En ese lapso, los estudiantes aprenden a identificar qué problemas requieren métodos numéricos para su solución y ven ejemplos de la propagación (o difusión) del error que puede ocurrir al aplicarlos. Además, reconocen cómo aproximar con precisión las soluciones de problemas que no se pueden resolver con exactitud y aprenden técnicas de estimación de cotas (o límites) del error en las aproximaciones. El resto del texto sirve como referencia para métodos no considerados en el curso. El tratamiento de un año o un semestre es consistente con los propósitos del texto.

Casi todos los conceptos del texto se ilustran mediante ejemplos; esta edición contiene más de 2,000 ejercicios probados en clase que abarcan desde aplicaciones elementales

de los métodos y los algoritmos hasta generalizaciones y extensiones de la teoría. Además, los conjuntos de ejercicios incluyen muchos problemas de aplicación de diversas áreas de la ingeniería, así como de las ciencias físicas, de la computación, biológicas y sociales. Las aplicaciones elegidas demuestran en forma concisa cómo se pueden aplicar los métodos numéricos en situaciones reales.

Existen varios paquetes de *software* para realizar cálculos matemáticos simbólicos. De éstos, predominan en el medio académico Derive, Maple y Mathematica. Las versiones escolares de estos paquetes están disponibles a precios razonables y funcionan en la mayoría de las computadoras. Aunque existen diferencias importantes entre los paquetes, tanto en desempeño como en precio, todos pueden realizar operaciones comunes de cálculo y álgebra.

El hecho de contar con un paquete para el cálculo simbólico puede ser muy útil en el estudio de las técnicas de aproximación. Los resultados de la mayor parte de nuestros ejemplos y ejercicios se generaron a partir de problemas para los que *pueden* determinarse los valores exactos, pues esto permite examinar el desempeño del método de aproximación. A menudo, las soluciones exactas se pueden obtener con relativa facilidad mediante cálculo simbólico. Además, para muchas técnicas numéricas, el análisis del error exige acotar una derivada ordinaria o parcial de orden superior de una función, lo cual puede ser una tarea tediosa y poco instructiva cuando se dominan las técnicas del cálculo. Las derivadas se pueden obtener rápidamente en forma simbólica y un poco de ingenio permite que un cálculo simbólico ayude también en el proceso de acotación.

Se eligió a Maple como paquete estándar debido al uso generalizado, pero Derive o Mathematica se pueden sustituir con sólo ligeras modificaciones. También, se agregaron ejemplos y ejercicios donde se tiene la impresión que un sistema de álgebra por computadora podría traer beneficios significativos y se analizaron los métodos de aproximación que usa Maple cuando *no* es posible resolver un problema de manera exacta.

Novedades en esta edición

La séptima edición incluye dos nuevas secciones importantes. El método del gradiente conjugado precondicionado se agregó al capítulo 7 para proporcionar un tratamiento más completo de la solución numérica de los sistemas de ecuaciones lineales. Se presenta como una técnica de aproximación iterativa para resolver sistemas lineales positivos definidos. De esta forma, es particularmente útil para aproximar la solución de sistemas dispersos de gran tamaño.

En el capítulo 10 se añadió una sección sobre métodos de homotopía y continuación. Éstos proporcionan una técnica muy distinta, que en fechas recientes ha llamado mucho la atención, para aproximar las soluciones de sistemas de ecuaciones no lineales.

También se añaden en todo el libro grandes listados de código Maple, pues los revisores consideraron útil esta característica en la sexta edición. Hemos actualizado todo el código Maple a la versión 6, que es la más reciente. Las personas familiarizadas con nuestras ediciones anteriores verán que casi todas las páginas mejoraron de alguna manera. Se actualizaron y revisaron las referencias y se han agregado nuevos ejercicios. Esperamos que todos estos cambios le parezcan benéficos para la enseñanza y el estudio del análisis numérico; la mayor parte de ellos han sido motivados por cambios en la presentación del material a nuestros propios estudiantes.

Otra modificación importante en esta edición es un sitio en Internet* en

http://www.as.ysu.edu/~faires/Numerical-Analysis/

En este sitio colocaremos programas actualizados conforme cambie el software, así como respuestas a los comentarios realizados por usuarios del libro. También podemos agregar nuevo material que podría incluirse en ediciones posteriores, en la forma de archivos PDF que pueden ser consultados por los usuarios.

Esperamos que esto amplíe la vida de la séptima edición, a la vez que mantenga actualizado el material.

*La información contenida en este sitio está en inglés. Asimismo, esta casa editorial no se hace responsable si en algún momento desaparece el sitio o cambia de dirección.

Algoritmos

Como en las ediciones anteriores, se proporciona un algoritmo detallado y estructurado sin el listado del programa para cada método en el texto. Los algoritmos aparecen de forma que los estudiantes puedan codificarlos, aun con poca experiencia en programación.

Los programas para cada algoritmo están escritos en FORTRAN, Pascal y C. Además, los hemos codificado por medio de Maple y Mathematica, así como MATLAB, un paquete de software ampliamente utilizado para aplicaciones del álgebra lineal. Esto debe garantizar que se dispone de un conjunto de programas para la mayor parte de los sistemas de cómputo.

Por medio de los algoritmos se obtienen programas que dan los resultados correctos para los ejemplos y ejercicios en el texto, pero de ningún modo se intentó escribir software profesional de carácter general. En particular, los algoritmos no siempre están escritos de una forma que conduce al programa *más efectivo* en términos de requisitos de tiempo o almacenamiento. Cuando ocurre un conflicto entre escribir un algoritmo extremadamente eficaz y uno algo distinto que ilustre mejor las características importantes del método, se opta por lo segundo.

Acerca de los complementos en la dirección www.thomsonlearning.com.mx

En el sitio, el lector encontrará información sobre este libro y podrá, además, consultar los archivos electrónicos de los algoritmos que aparecen en el texto (en distintos formatos). Para cada algoritmo hay un programa C, Fortran, Maple, Mathematica, MATLAB y Pascal; y para cada sistema hay varios programas, cuya aplicación depende de la versión del software que se emplee; esos programas se ejemplifican con un problema del texto, de modo que el usuario pueda resolverlo en el lenguaje de su elección e identifique la entrada (INPUT) y la salida (OUTPUT); éstos pueden también modificarse para resolver otros problemas. Las entradas y salidas son casi las mismas en cada sistema de programación.

Los programas pueden correrse en una computadora que posea los sistemas operativos MS-DOS, Windows o Macintosh. Sin embargo, se requiere un software apropiado, como un compilador para Pascal, Fortran, C, o algún sistema algebraico para computadora (Maple, Mathematica o MATLAB). El lector encontrará seis subdirectorios para cada lenguaje y los archivos complementarios.

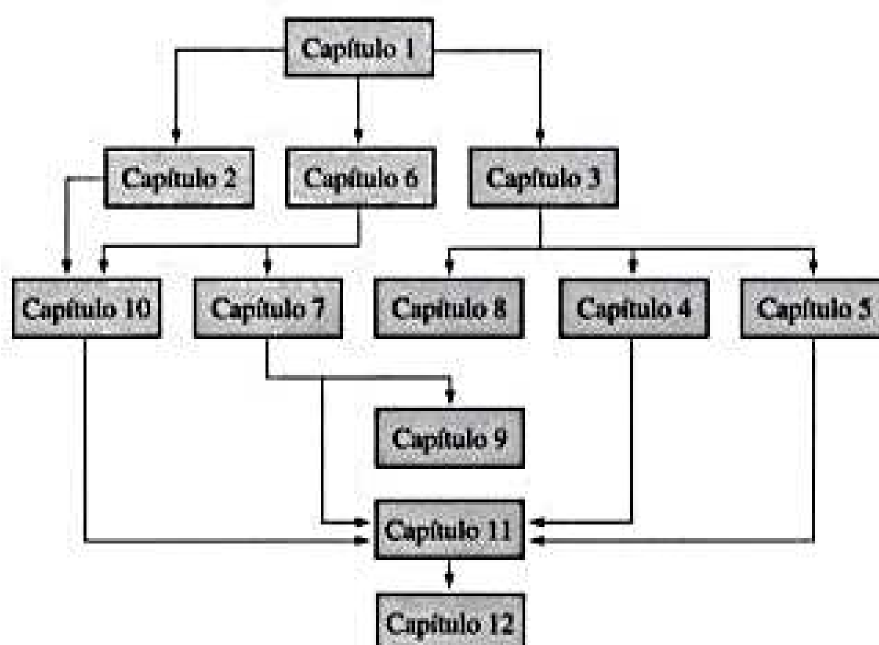
Todos los programas están en archivos ASCII y hojas de cálculo; y pueden modificarse mediante un procesador de palabras, capaz de crear un archivo estándar de ASCII (de los llamados "sólo texto").

Los archivos README se presentan en formato ASCII y PDF, y se incluyen con los archivos del programas, de manera que los sistemas de programación puedan ejecutarse en forma individual.

Sugerencias para un curso

Análisis numérico está diseñado para que los profesores puedan elegir entre los temas, así como el nivel de rigor teórico y el énfasis en las aplicaciones. En concordancia con estos propósitos, proporcionamos referencias detalladas para los resultados no demostrados en el texto y las aplicaciones utilizadas para indicar la importancia práctica de los métodos. Las referencias son las que tienen más posibilidades de ser halladas en las bibliotecas de las universidades y se actualizaron para reflejar la edición más reciente en el momento en que este libro se imprimió. También incluimos citas de artículos originales de investigación cuando consideramos que el material es accesible a nuestros lectores.

En el siguiente diagrama de flujo se indican los prerrequisitos de cada capítulo. La única desviación de este diagrama se describe en la nota al pie de página, al inicio de la sección 3.4. La mayor parte de las secuencias posibles que pueden generarse con este diagrama, los autores las utilizaron en *Youngstown State University*.



Agradecimientos

Nos sentimos muy afortunados porque muchos de nuestros estudiantes y colegas nos han comunicado sus impresiones acerca de las ediciones anteriores de este libro. Todos estos comentarios fueron tomados en cuenta y se procuró incluir todas las sugerencias acordes con los principios del libro; asimismo, agradecemos a todos aquellos que se han tomado un tiempo para contactarnos e informarnos de mejoras que podemos hacer en versiones posteriores.

En particular, queremos agradecer el esfuerzo de las siguientes personas:

Glen Granzow, Idaho State University
José Miguel, Universidad Peruana Cayetano Heredia, Lima, Perú
John M. Neuberger, Northern Arizona University
L. G. de Pillis, Harvey Mudd College

Agradecemos en particular a nuestro amigo y alguna vez alumno, Jim Baglama, de *Ball State University*. Jim estuvo de acuerdo en revisar ampliamente esta edición y nos ayudó a actualizar las secciones de bibliografía y software. Es muy gratificante ver el desarrollo de nuestros estudiantes en su profesión.

Otra persona que se desempeña bien en su profesión, aunque de manera muy distinta, es nuestro editor Gary Ostedt, gerente sobresaliente de nuestros proyectos y gran amigo en lo personal. Extrañaremos en gran medida su dirección y apoyo, y aprovechamos la oportunidad para desearle lo mejor en su pronto retiro de Brooks/Cole.

Igual que en las ediciones anteriores de este libro, aprovechamos la ayuda de los estudiantes de Youngstown State University para preparar la séptima edición. Nuestra hábil asistente para esta edición fue Laurie Marinelli, a la que agradecemos su trabajo. También queremos expresar nuestra gratitud a los colegas académicos y administrativos de Youngstown State University por darnos la oportunidad y facilitarnos los medios para concluir este proyecto.

Por último, quisiéramos agradecer a quienes han utilizado y adoptado las diversas ediciones de *Análisis numérico* en estos años. Es maravilloso saber de tantos estudiantes y jóvenes profesores que utilizan nuestro libro en su primer encuentro con el estudio de los métodos numéricos. Esperamos que esta edición continúe esta tendencia y apoye el gusto de los estudiantes por el análisis numérico. Si usted tiene sugerencias para mejorar el material que puedan incorporarse en las próximas ediciones del libro, agradeceremos sus comentarios en las siguientes direcciones de correo electrónico:

Richard L. Burden
burden@math.ysu.edu
J. Douglas Faires
faires@math.ysu.edu

CAPÍTULO

1

Preliminares matemáticos

. . .

En los primeros cursos de química se introduce la *ley del gas ideal*,

$$PV = NRT,$$

que relaciona la presión P, el volumen V, la temperatura T y el número de moles N de un gas "ideal". En esta ecuación, R es una constante que depende del sistema de medida.

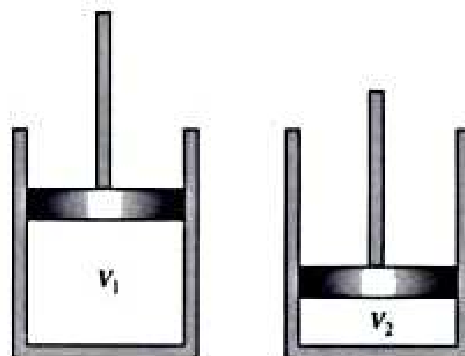
Suponga que se realizan dos experimentos para comprobar esta ley con el mismo gas en cada caso. En el primer experimento,

$$P = 1.00 \text{ atm}, \qquad V = 0.100 \text{ m}^3,$$
$$N = 0.00420 \text{ mol}, \qquad R = 0.08206.$$

Por medio de la ley del gas ideal se predice que la temperatura del gas será

$$T = \frac{PV}{NR} = \frac{(1.00)(0.100)}{(0.00420)(0.08206)} = 290.15 \text{ K} = 17 \text{ °C}.$$

Al medir la temperatura del gas, vemos que la temperatura real es 15 °C.



Luego, repetimos el experimento con los mismos valores de R y N, pero se incrementa la presión por un factor de dos y se reduce el volumen por el mismo factor. Como el producto PV es el mismo, la temperatura prevista aún es 17 °C, pero ahora la temperatura real del gas es de 19 °C.

Por supuesto que se sospecha de la ley del gas ideal, pero antes de concluir que la ley no es válida en esta situación es necesario examinar los datos para ver si el error se puede atribuir a los resultados experimentales. En caso afirmativo, podríamos determinar la precisión necesaria en nuestros resultados experimentales para garantizar que no ocurra un error de esta magnitud.

El análisis del error que surge en los cálculos es un tema importante en el análisis numérico y se presenta en la sección 1.2. Esta aplicación particular se considera en el ejercicio 28 de esa sección.

Este capítulo contiene un breve repaso de temas de cálculo elemental de una variable, necesarios en capítulos posteriores, junto con una introducción a la convergencia, el análisis del error y la representación de los números en los dispositivos utilizados para la realización de los cálculos.

1.1 Repaso de cálculo

Los conceptos de *límite* y *continuidad* de una función son fundamentales en el estudio del cálculo diferencial.

Definición 1.1 Una función f definida en un conjunto X de números reales tiene el **límite** L en x_0, denotado por

$$\lim_{x \to x_0} f(x) = L,$$

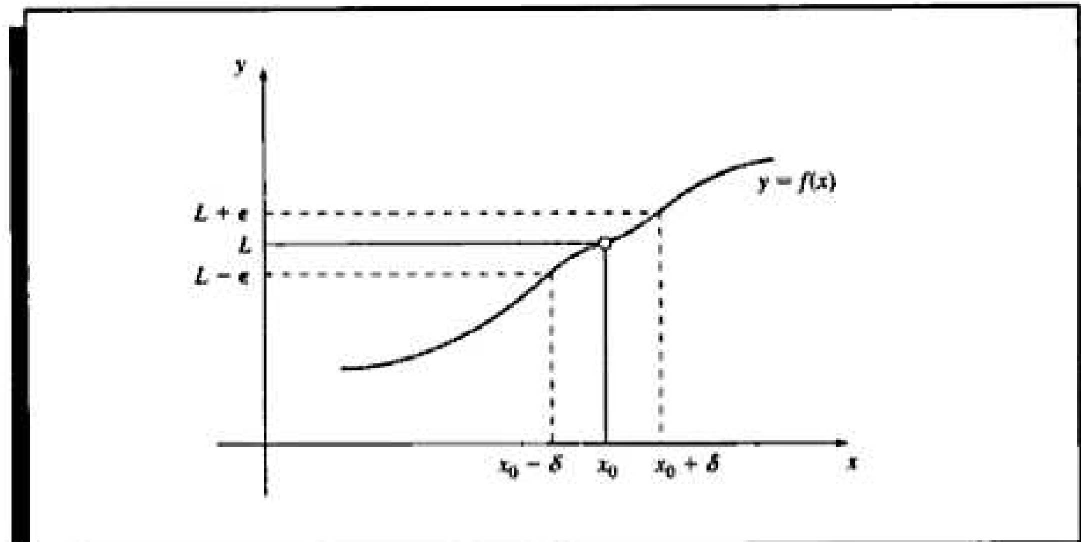
si, dado cualquier número real $\epsilon > 0$, existe un número real $\delta > 0$ tal que $|f(x) - L| < \epsilon$, siempre que $x \in X$ y $0 < |x - x_0| < \delta$. (Véase la figura 1.1.) ■

Definición 1.2 Sea f una función definida en un conjunto X de números reales y $x_0 \in X$. Entonces f es **continua** en x_0 si

$$\lim_{x \to x_0} f(x) = f(x_0).$$

La función f es continua en el conjunto X si es continua en cada número en X. ■

Figura 1.1



$C(X)$ denota al conjunto de funciones que son continuas en X. Cuando X es un intervalo de la recta real, se omiten los paréntesis en esta notación. Por ejemplo, el conjunto de todas las funciones continuas en el intervalo cerrado $[a,b]$ se denota $C[a,b]$.

El *límite de una sucesión* de números reales o complejos se define de manera similar.

Definición 1.3 Sea $\{x_n\}_{n=1}^{\infty}$ una sucesión infinita de números reales o complejos. La sucesión $\{x_n\}_{n=1}^{\infty}$ tiene el **límite** x (**converge a** x) si, para cualquier $\epsilon > 0$, existe un entero positivo $N(\epsilon)$ tal que $|x_n - x| < \epsilon$, siempre que $n > N(\epsilon)$. La notación

$$\lim_{n\to\infty} x_n = x, \qquad \text{o} \qquad x_n \to x \qquad \text{cuando} \qquad n \to \infty,$$

significa que la sucesión $\{x_n\}_{n=1}^{\infty}$ converge a x. ■

En el siguiente teorema se relacionan los conceptos de convergencia y continuidad.

Teorema 1.4 Si f es una función definida en un conjunto X de números reales y $x_0 \in X$, entonces las siguientes afirmaciones son equivalentes:

a. f es continua en x_0;

b. Si $\{x_n\}_{n=1}^{\infty}$ es cualquier sucesión en X que converge a x_0, entonces $\lim_{n\to\infty} f(x_n) = f(x_0)$. ■

Supondremos que son continuas las funciones por considerar en el análisis de los métodos numéricos, pues éste es un requisito mínimo para tener un comportamiento predecible. Las funciones discontinuas pueden interrumpirse en los puntos de interés, lo que puede causar dificultades al intentar aproximar una solución a un problema. Por lo general los supuestos más elaborados acerca de una función conducen a mejores resultados de aproximación. Por ejemplo, una función con gráfica uniforme se comportará, por lo gene-

ral, de manera más predecible que una en forma de sierra. La condición de suavidad se basa en el concepto de derivada.

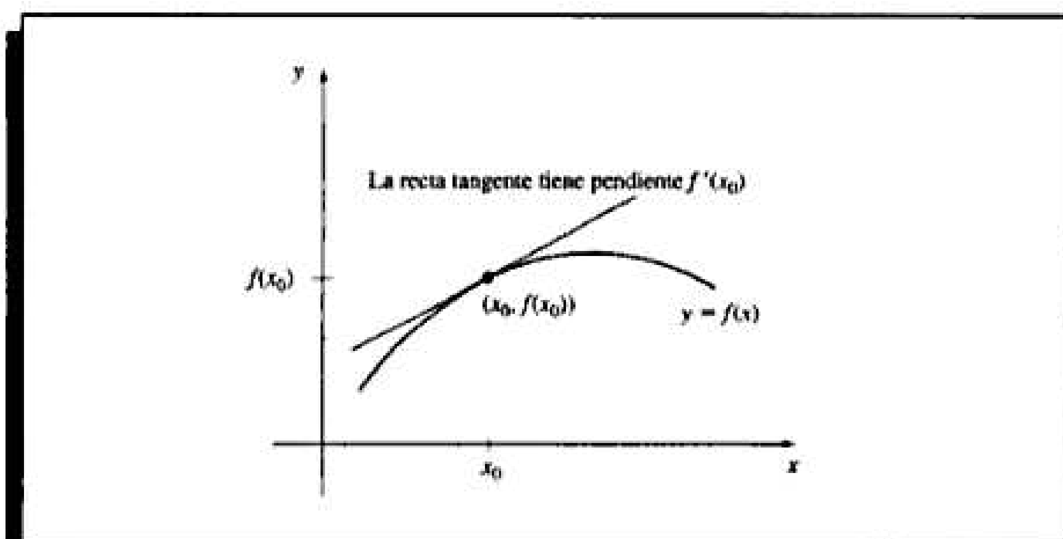
Definición 1.5 Sea f una función definida en un intervalo abierto que contiene a x_0. La función f es **derivable en** x_0 si

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

existe. El número $f'(x_0)$ es la **derivada** de f en x_0. Una función que tiene derivada en cada número de un conjunto X es **derivable** en X.

La derivada de f en x_0 es la pendiente de la recta tangente a la gráfica de f en $(x_0, f(x_0))$, como se muestra en la figura 1.2. ■

Figura 1.2



Teorema 1.6 Si la función f es derivable en x_0, entonces f es continua en x_0. ■

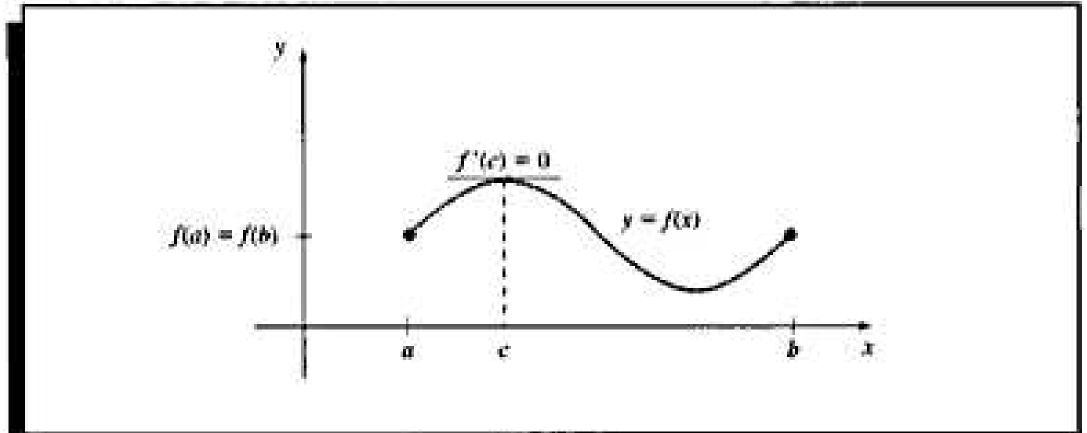
El conjunto de todas las funciones que tienen n derivadas continuas en X se denota $C^n(X)$, y el conjunto de funciones que tienen derivadas de todos los órdenes en X se denota $C^\infty(X)$. Las funciones polinomiales, racionales, trigonométricas, exponenciales y logarítmicas están en $C^\infty(X)$, donde X consta de todos los números para los que están definidas las funciones. Si X es un intervalo de la recta real, se omiten de nuevo los paréntesis en esta notación.

Los siguientes teoremas son fundamentales en la deducción de métodos para la estimación del error. Las demostraciones de estos teoremas y de los demás resultados sin referencia en esta sección se pueden encontrar en cualquier texto de cálculo.

Teorema 1.7 **(Teorema de Rolle)**

Suponga que $f \in C[a,b]$ y que f es derivable en (a, b). Si $f(a) = f(b)$, entonces existe un número c en (a, b) tal que $f'(c) = 0$. (Véase la figura 1.3.) ■

Figura 1.3

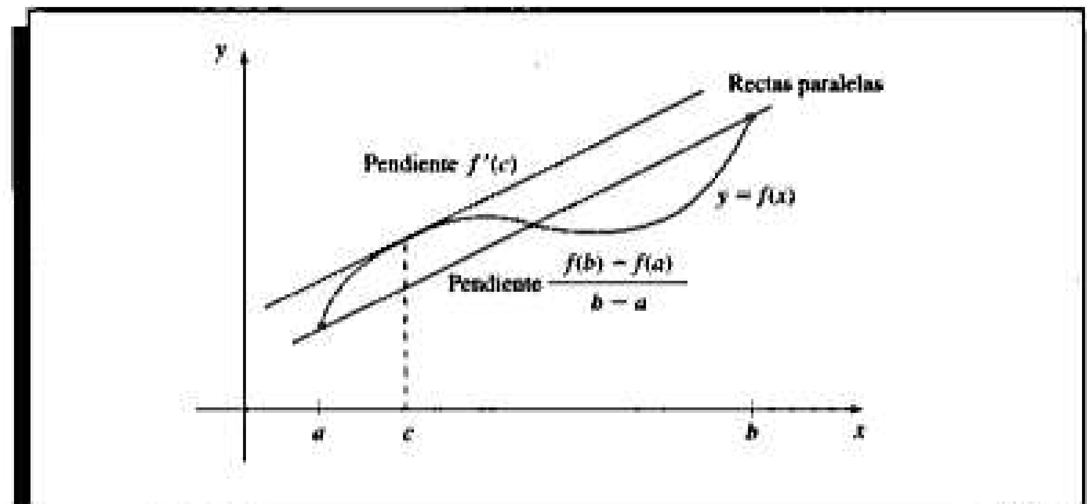


Teorema 1.8 **(Teorema del valor medio)**

Si $f \in C[a, b]$ y f es derivable en (a, b), entonces existe un número c en (a, b) tal que

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \quad \text{(Véase la figura 1.4.)}$$ ■

Figura 1.4

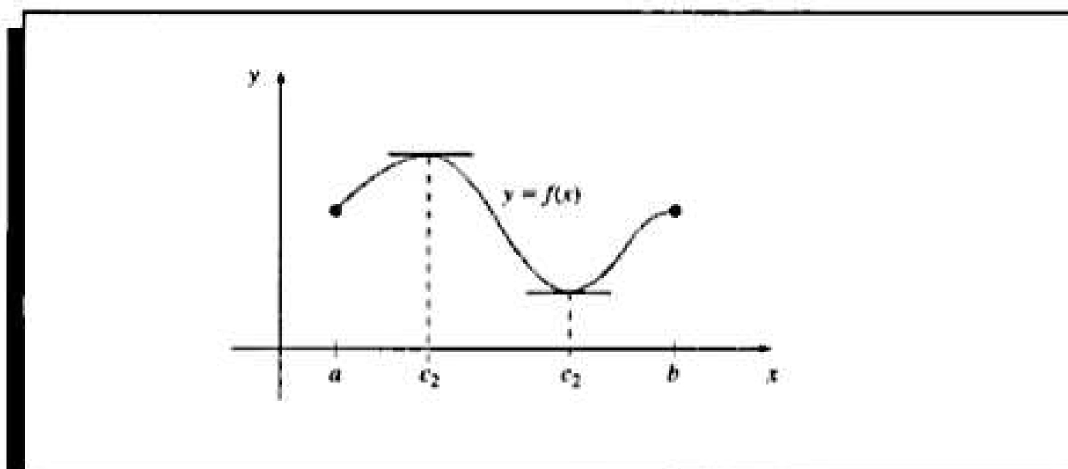


Teorema 1.9 **(Teorema de los valores extremos)**

Si $f \in C[a, b]$, entonces existen $c_1, c_2 \in [a, b]$ tales que $f(c_1) \leq f(x) \leq f(c_2)$ para toda $x \in [a, b]$. Además, si f es derivable en (a, b), entonces los números c_1 y c_2 aparecen en los extremos de $[a, b]$, o bien donde se anula f'. (Véase la figura 1.5.) ■

Como se mencionó en el prefacio, cuando sea adecuado usaremos el sistema de álgebra por computadora, Maple. Los sistemas de álgebra por computadora son útiles para la derivación simbólica y el trazo de gráficas. Ambas técnicas se ilustran en el ejemplo 1.

Figura 1.5



EJEMPLO 1 Determine $\max_{a \le x \le b} |f(x)|$ para

$$f(x) = 5 \cos 2x - 2x \operatorname{sen} 2x$$

en los intervalos [1, 2] y [0.5, 1].

Primero ilustraremos las capacidades gráficas de Maple. Para tener acceso al paquete de graficación se escribe la instrucción

```
>with(plots);
```

A continuación aparecen las instrucciones del paquete. Se introduce la definición de f al escribir

```
>f:=5*cos(2*x)-2*x*sen(2*x);
```

La respuesta de Maple es

$$f := 5 \cos(2x) - 2x \operatorname{sen}(2x)$$

Para graficar f en el intervalo [0.5, 2], use la instrucción

```
>plot(f,x=0.5..2);
```

La gráfica aparece como en la figura 1.6, en la cual se pueden determinar las coordenadas de cualquier punto sobre la gráfica al mover el apuntador del ratón al punto deseado y oprimir el botón izquierdo. Esta técnica se utiliza para estimar las intersecciones con los ejes y los extremos de las funciones.

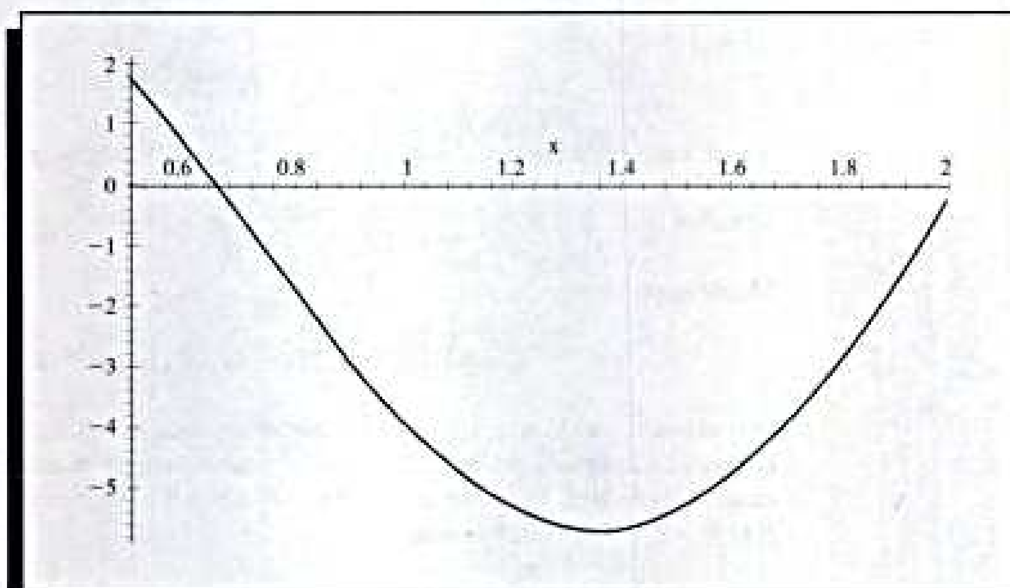
Completamos el ejemplo usando el teorema de valores extremos. Primero, consideremos el intervalo [1, 2]. Para obtener la primera derivada $g = f'$, se escribe

```
>g:=diff(f,x);
```

la respuesta de Maple es

$$g: = -12 \operatorname{sen}(2x) - 4x \cos(2x)$$

Figura 1.6



Luego, podemos resolver $g(x) = 0$ para $1 \leq x \leq 2$, con la instrucción

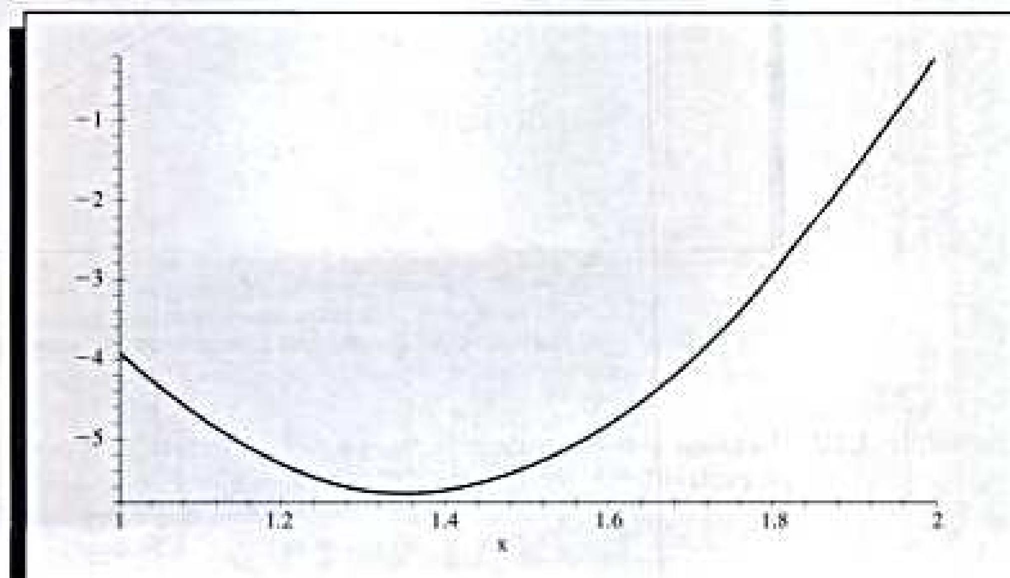
```
>fsolve(g,x,1..2);
```

con lo cual se obtiene 1.358229874, y se calcula $f(1.358229874)$ mediante

```
>evalf(subs(x=1.358229874,f));
```

Como $f(1) = -3.899329037$ y $f(2) = -0.241008124$, tenemos, para el intervalo [1, 2], un valor máximo de $f(2) = -0.241008124$, como se muestra en la figura 1.7, y un

Figura 1.7



valor mínimo aproximado de $f(1.358229874) = -5.675301338$. Por tanto,

$$\max_{1 \le x \le 2} |5 \cos 2x - 2x \operatorname{sen} 2x| \approx |f(1.358229874)| = 5.675301338.$$

Si tratamos de resolver $g(x) = 0$, para $0.5 \le x \le 1$, vemos que al introducir

```
>fsolve(g,x,0.5..1);
```

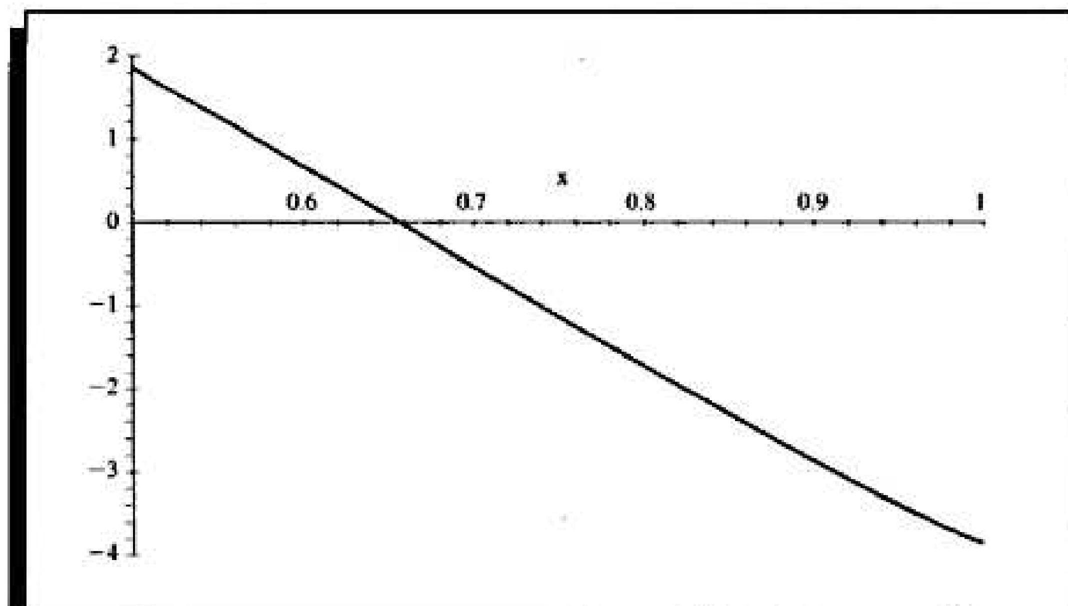
Maple responde con

$$\text{fsolve}(-12 \operatorname{sen}(2x) - 4x \cos(2x), x, .5..1)$$

lo cual indica que Maple no pudo encontrar una solución en [0.5, 1]. Si grafica g, verá que no hay soluciones en este intervalo, y el máximo ocurre en un extremo. Por tanto, f' nunca se anula en [0.5, 1], como se muestra en la figura 1.8, y como $f(0.5) = 1.860040545$ y $f(1) = -3.899329037$, tenemos

$$\max_{0.5 \le x \le 1} |5 \cos 2x - 2x \operatorname{sen} 2x| \approx |f(1)| = 3.899329037. \qquad \blacksquare$$

Figura 1.8



El otro concepto básico del cálculo que usaremos ampliamente es la integral de Riemann.

Definición 1.10 **La integral de Riemann** de la función f en el intervalo $[a, b]$ es el siguiente límite, si éste existe:

$$\int_a^b f(x)\,dx = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(z_i)\,\Delta x_i,$$

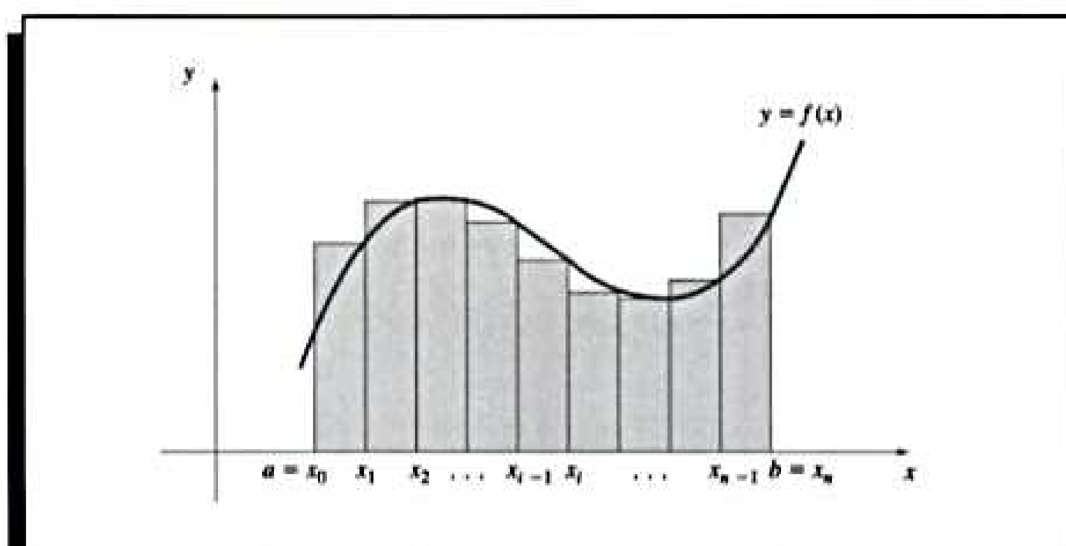
donde los números $x_0, x_1, \ldots, x_n$ satisfacen $a = x_0 \leq x_1 \leq \cdots \leq x_n = b$, y $\Delta x_i = x_i - x_{i-1}$ para toda $i = 1, 2, \ldots, n$ y z_i se elige de manera arbitraria en el intervalo $[x_{i-1}, x_i]$. ■

Toda función continua f en $[a, b]$ es Riemann-integrable en $[a, b]$. Esto nos permite elegir, para fines de cálculo, los puntos x_i uniformemente distribuidos en $[a, b]$, y para cada $i = 1, 2, \ldots, n$, elegir $z_i = x_i$. En este caso,

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i),$$

donde los números que aparecen en la figura 1.9 como x_i son $x_i = a + i(b-a)/n$.

Figura 1.9



En nuestro estudio del análisis numérico necesitaremos otros dos resultados. El primero es una generalización del teorema del valor medio para integrales.

Teorema 1.11 **(Teorema del valor medio ponderado para integrales)**
Suponga que $f \in C[a, b]$, que la integral de Riemann de g existe en $[a, b]$ y que $g(x)$ no cambia de signo en $[a, b]$. Entonces existe un número c en (a, b) tal que

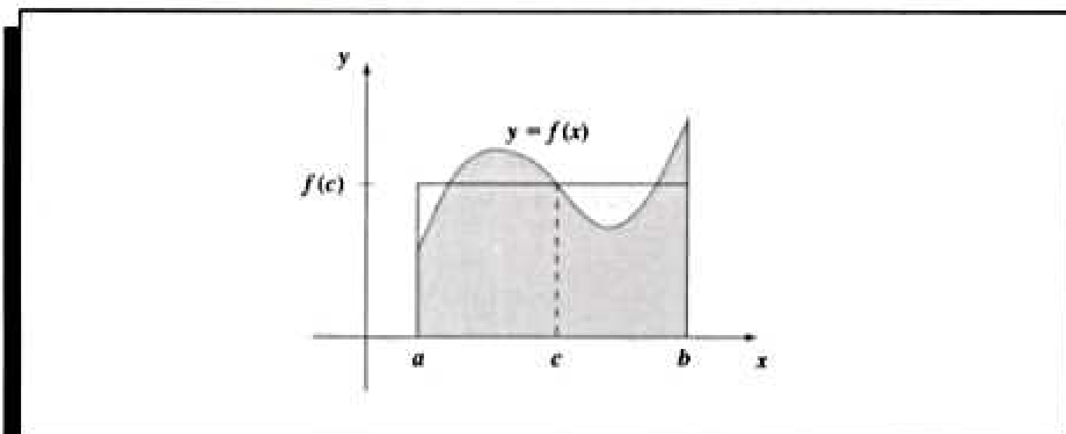
$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx.$$ ■

Cuando $g(x) \equiv 1$, el teorema 1.11 es el del valor medio para integrales que proporciona el **valor promedio** de la función f en el intervalo $[a, b]$ como

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

(Véase la figura 1.10.)

Figura 1.10



Por lo general, la demostración del teorema 1.11 no se da en un curso básico de cálculo, pero se puede encontrar en la mayor parte de los textos de análisis (véase, por ejemplo, [Fu, p. 162]).

El otro teorema que necesitaremos y que normalmente no se presenta en un curso básico de cálculo se deduce al aplicar de manera sucesiva el teorema de Rolle a $f, f', \ldots,$ y, por último, a $f^{(n-1)}$.

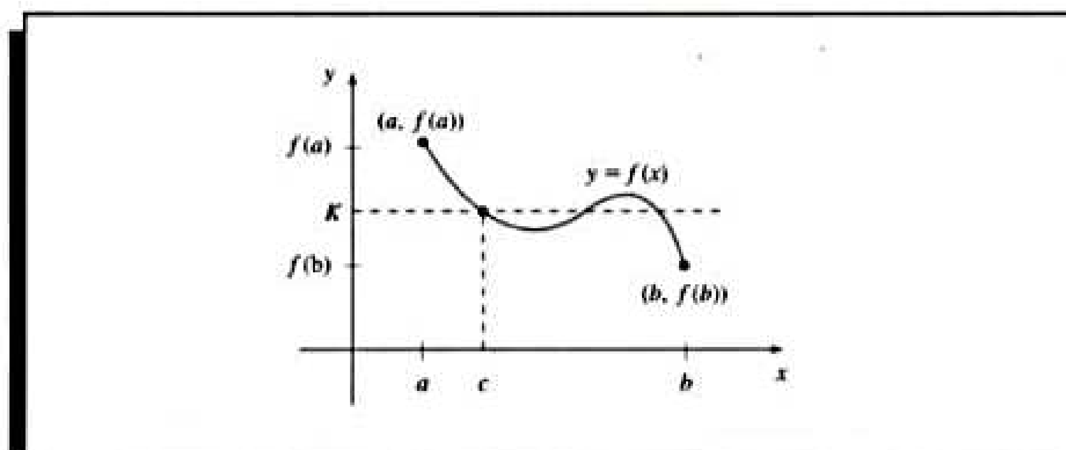
Teorema 1.12 **(Teorema generalizado de Rolle)**
Suponga que $f \in C[a, b]$ es n veces derivable en (a, b). Si $f(x)$ se anula en los $n + 1$ números distintos $x_0, \ldots, x_n$ en $[a, b]$, entonces existe un número c en (a, b) tal que $f^{(n)}(c) = 0$. ■

El siguiente teorema es el del valor intermedio. Aunque su enunciado parece razonable, la demostración está fuera del alcance de un curso usual de cálculo. Sin embargo, se puede encontrar en la mayor parte de los textos de análisis (véase, por ejemplo, [Fu, p. 67]).

Teorema 1.13 **(Teorema del valor intermedio)**
Si $f \in C[a, b]$ y K es cualquier número entre $f(a)$ y $f(b)$, entonces existe un número c en (a, b) tal que $f(c) = K$. ■

En la figura 1.11 se muestra una elección del número garantizado por el teorema del valor intermedio. En este ejemplo hay otras dos posibilidades.

Figura 1.11



EJEMPLO 2 Para mostrar que $x^5 - 2x^3 + 3x^2 - 1 = 0$ tiene una solución en el intervalo [0, 1], considere $f(x) = x^5 - 2x^3 + 3x^2 - 1$. Puesto que

$$f(0) = -1 < 0 < 1 = f(1)$$

y f es continua, el teorema del valor intermedio indica que existe un número x en $0 < x < 1$ para el que $x^5 - 2x^3 + 3x^2 - 1 = 0$. ■

Como vimos en el ejemplo 2, el teorema del valor intermedio sirve para determinar si existen soluciones de ciertos problemas. Sin embargo, no proporciona un método eficaz para determinar tales soluciones. Estudiaremos este tema en el capítulo 2.

El último teorema en este repaso de cálculo describe los polinomios de Taylor. Éstos se usan ampliamente en el análisis numérico.

Teorema 1.14 **(Teorema de Taylor)**

Suponga que $f \in C^n[a, b]$, que $f^{(n+1)}$ existe en $[a, b]$ y $x_0 \in [a, b]$. Para cada $x \in [a, b]$, existe un número $\xi(x)$ entre x_0 y x tal que

$$f(x) = P_n(x) + R_n(x),$$

donde

$$\begin{aligned} P_n(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \end{aligned}$$

y

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}.$$ ■

En este caso, $P_n(x)$ es el ***n*-ésimo polinomio de Taylor** para f respecto a x_0, y $R_n(x)$ se llama el **término del residuo** (o **error de truncamiento**) asociado a $P_n(x)$. La serie infinita obtenida al tomar el límite de $P_n(x)$ cuando $n \to \infty$ es la **serie de Taylor** para f en torno a x_0. En el caso $x_0 = 0$, el polinomio de Taylor suele llamarse **polinomio de Maclaurin**, y la serie de Taylor se nombra **serie de Maclaurin**.

El término **error de truncamiento** se refiere al error implícito al usar una suma truncada, o finita, para aproximar la suma de una serie infinita.

EJEMPLO 3 Determine (a) el segundo y (b) el tercer polinomio de Taylor para $f(x) = \cos x$ respecto a $x_0 = 0$, y use estos polinomios para aproximar cos(0.01). (c) Con el tercer polinomio de Taylor y su término de residuo aproxime $\int_0^{0.1} \cos x\, dx$.

Como $f \in C^\infty(\mathbb{R})$, el teorema de Taylor se puede aplicar a cualquier $n \geq 0$. Además,

$$f'(x) = -\operatorname{sen} x, \quad f''(x) = -\cos x, \quad f'''(x) = \operatorname{sen} x \quad \text{y} \quad f^{(4)}(x) = \cos x.$$

de modo que

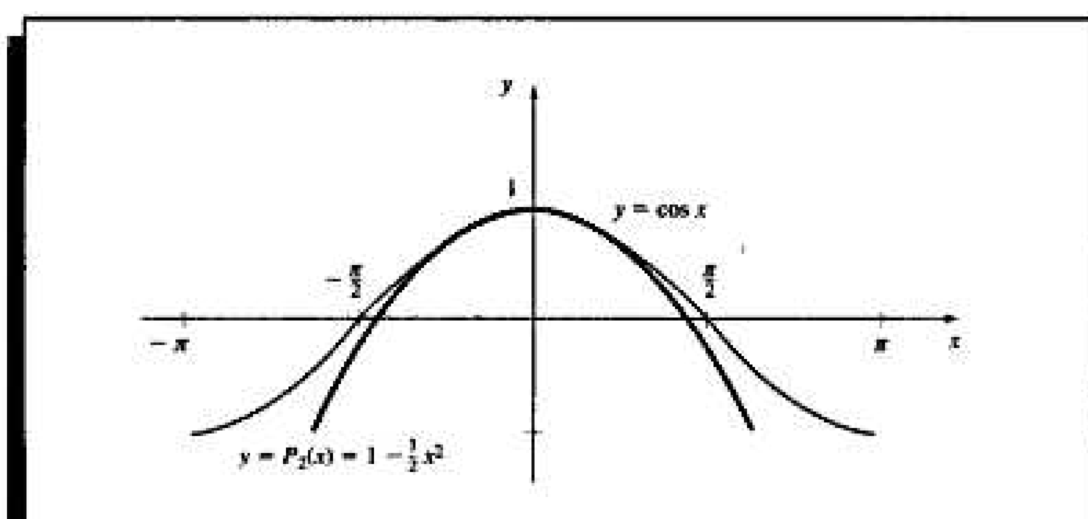
$$f(0) = 1, \quad f'(0) = 0, \quad f''(0) = -1 \qquad \text{y} \qquad f'''(0) = 0.$$

a. Para $n = 2$ y $x_0 = 0$, tenemos

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{6}x^3 \operatorname{sen} \xi(x),$$

donde $\xi(x)$ es un número entre 0 y x. (Véase la figura 1.12.)

Figura 1.12



Para $x = 0.01$, el polinomio de Taylor y el término del residuo son

$$\cos 0.01 = 1 - \frac{1}{2}(0.01)^2 + \frac{1}{6}(0.01)^3 \operatorname{sen} \xi(x)$$
$$= 0.99995 + 0.1\overline{6} \times 10^{-6} \operatorname{sen} \xi(x),$$

donde $0 < \xi(x) < 0.01$. (La barra sobre el 6 en $0.1\overline{6}$ se usa para indicar que este dígito se repite de manera indefinida.) Puesto que $|\operatorname{sen} \xi(x)| < 1$ para toda x, tenemos

$$|\cos 0.01 - 0.99995| \leq 0.1\overline{6} \times 10^{-6},$$

de modo que la aproximación 0.99995 coincide por lo menos con los primeros cinco dígitos de cos 0.01, y

$$0.9999483 < 0.99995 - 1.\overline{6} \times 10^{-6} \leq \cos 0.01 \leq 0.99995 + 1.\overline{6} \times 10^{-6}$$
$$< 0.9999517.$$

La cota de error es mucho mayor que el error real. Esto se debe, en parte, a la pobre cota que usamos para $|\operatorname{sen} \xi(x)|$. Se puede demostrar que para todo valor de x, tenemos $|\operatorname{sen} x| \leq |x|$. Como $0 \leq \xi < 0.01$, podríamos usar el hecho de que $|\operatorname{sen} \xi(x)| \leq 0.01$ en la fórmula para el error, lo que produce la cota $0.1\overline{6} \times 10^{-8}$.

b. Como $f'''(0) = 0$, el tercer polinomio de Taylor con término de residuo en torno a $x_0 = 0$ es

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 \cos \tilde{\xi}(x),$$

donde $0 < \tilde{\xi}(x) < 0.01$. El polinomio de aproximación es el mismo, y la aproximación aún es 0.99995, pero ahora tenemos una mucho mejor garantía de precisión. Puesto que $|\cos \tilde{\xi}(x)| \le 1$ para toda x, tenemos

$$\left|\frac{1}{24}x^4 \cos \tilde{\xi}(x)\right| \le \frac{1}{24}(0.01)^4(1) \approx 4.2 \times 10^{-10}.$$

De modo que

$$|\cos 0.01 - 0.99995| \le 4.2 \times 10^{-10},$$

y

$$\begin{aligned} 0.99994999958 = 0.99995 - 4.2 \times 10^{-10} \\ \le \cos 0.01 \le 0.99995 + 4.2 \times 10^{-10} = 0.99995000042. \end{aligned}$$

En las dos primeras partes de este ejemplo se ilustran los dos objetivos del análisis numérico. El primero es encontrar una aproximación, que los polinomios de Taylor proporcionan en ambas partes. El segundo es determinar la precisión de la aproximación. En este caso, el tercer polinomio de Taylor fue mucho más informativo que el segundo, aunque ambos dieron la misma aproximación.

c. Usamos el tercer polinomio de Taylor para obtener

$$\begin{aligned} \int_0^{0.1} \cos x\, dx &= \int_0^{0.1} \left(1 - \frac{1}{2}x^2\right) dx + \frac{1}{24}\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx \\ &= \left[x - \frac{1}{6}x^3\right]_0^{0.1} + \frac{1}{24}\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx \\ &= 0.1 - \frac{1}{6}(0.1)^3 + \frac{1}{24}\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx. \end{aligned}$$

Por tanto,

$$\int_0^{0.1} \cos x\, dx \approx 0.1 - \frac{1}{6}(0.1)^3 = 0.0998\bar{3}.$$

Se puede determinar una cota para el error en esta aproximación a partir de la integral del término del residuo de Taylor y el hecho de que $|\cos \tilde{\xi}(x)| \le 1$ para toda x:

$$\begin{aligned} \frac{1}{24}\left|\int_0^{0.1} x^4 \cos \tilde{\xi}(x)\, dx\right| &\le \frac{1}{24}\int_0^{0.1} x^4 |\cos \tilde{\xi}(x)|\, dx \\ &\le \frac{1}{24}\int_0^{0.1} x^4\, dx = 8.\bar{3} \times 10^{-8}. \end{aligned}$$

Como el valor real de esta integral es

$$\int_0^{0.1} \cos x\, dx = \operatorname{sen} x \Big]_0^{0.1} = \operatorname{sen} 0.1 \approx 0.099833417,$$

el error real para esta aproximación es 8.332×10^{-8}, que está dentro de la cota de error. ■

En el ejemplo 3, también podríamos usar un sistema de álgebra por computadora. Por medio de Maple, se define a f como

```
>f:=cos(x);
```

Maple nos permite colocar varios enunciados en un renglón y usar los dos puntos para eliminar las respuestas de Maple. Por ejemplo, obtenemos el tercer polinomio de Taylor con

```
>s3:=taylor(f,x=0,4): p3:=convert(s3, polynom);
```

La primera parte calcula la serie de Taylor con cuatro términos (grado 3) y el residuo desarrollado respecto a $x_0 = 0$. La segunda parte convierte la serie `s3` en el polinomio `p3` al eliminar el residuo. Para obtener 11 cifras decimales en el resultado, introducimos

```
>Digits:=11;
```

y evaluamos $f(0.01)$, $P_3(0.01)$ y $|f(0.01) - P_3(0.01)|$ con

```
>y1:=evalf(subs(x=0.01,f));
>y2:=evalf(subs(x=0.01,p3));
>err:=abs(y1-y2);
```

Esto produce $y_1 = f(0.01) = 0.99995000042$, $y_2 = P_3(0.01) = 0.99995000000$ y $|f(0.01) - P_3(0.01)| = .42 \times 10^{-9}$.

Para obtener una gráfica similar a la figura 1.12, introducimos

```
>plot({ f,p3 },x=-Pi..Pi);
```

Las instrucciones para las integrales son

```
>q1:=int(f, x=0..0.1);
>q2:=int(p3, x=0..0.1);
>err:=abs(q1-q2);
```

con lo cual se obtienen los valores

$$q_1 = \int_0^{0.1} f(x)\, dx = 0.099833416647 \quad \text{y} \quad q_2 = \int_0^{0.1} P_3(x)\, dx = 0.099833333333,$$

con error 0.83314×10^{-7}.

En los incisos (a) y (b) del ejemplo se muestra cómo dos técnicas pueden producir la misma aproximación pero con diversas garantías de precisión. Recuerde que determinar las aproximaciones es sólo parte de nuestro objetivo. Una parte igualmente importante es determinar al menos una cota para la precisión de la aproximación.

CONJUNTO DE EJERCICIOS 1.1

1. Demuestre que las siguientes ecuaciones tienen al menos una solución en los intervalos dados.

a. $x \cos x - 2x^2 + 3x - 1 = 0$, $[0.2, 0.3]$ y $[1.2, 1.3]$

b. $(x-2)^2 - \ln x = 0$, $[1, 2]$ y $[e, 4]$

c. $2x \cos(2x) - (x-2)^2 = 0$, $[2, 3]$ y $[3, 4]$

d. $x - (\ln x)^x = 0$, $[4, 5]$

2. Determine intervalos que contengan soluciones a las siguientes ecuaciones.

a. $x - 3^{-x} = 0$

b. $4x^2 - e^x = 0$

c. $x^3 - 2x^2 - 4x + 3 = 0$

d. $x^3 + 4.001x^2 + 4.002x + 1.101 = 0$

3. Demuestre que $f'(x)$ se anula al menos una vez en los intervalos dados.

a. $f(x) = 1 - e^x + (e-1)\operatorname{sen}((\pi/2)x)$, $[0,1]$

b. $f(x) = (x-1)\tan x + x \operatorname{sen} \pi x$, $[0, 1]$

c. $f(x) = x \operatorname{sen} \pi x - (x-2)\ln x$, $[1, 2]$

d. $f(x) = (x-2)\operatorname{sen} x \ln(x+2)$, $[-1, 3]$

4. Determine $\max_{a \le x \le b} |f(x)|$ para las siguientes funciones e intervalos.

a. $f(x) = (2 - e^x + 2x)/3$, $[0, 1]$

b. $f(x) = (4x-3)/(x^2-2x)$, $[0.5, 1]$

c. $f(x) = 2x\cos(2x) - (x-2)^2$, $[2, 4]$

d. $f(x) = 1 + e^{-\cos(x-1)}$, $[1, 2]$

5. Use el teorema del valor intermedio y el teorema de Rolle para mostrar que la gráfica de $f(x) = x^3 + 2x + k$ cruza el eje x exactamente una vez, sin importar el valor de la constante k.

6. Suponga que $f \in C[a,b]$ y que $f'(x)$ existe en (a, b). Demuestre que si $f'(x) \neq 0$ para toda x en (a, b), entonces puede existir a lo sumo un número p en $[a, b]$ tal que $f(p) = 0$.

7. Sea $f(x) = x^3$.

a. Determine el segundo polinomio de Taylor $P_2(x)$ en torno a $x_0 = 0$.

b. Calcule $R_2(0.5)$ y el error real al usar $P_2(0.5)$ para aproximar $f(0.5)$.

c. Repita el inciso (a) usando $x_0 = 1$.

d. Repita el inciso (b) con el polinomio del inciso (c).

8. Obtenga el tercer polinomio de Taylor $P_3(x)$ para la función $f(x) = \sqrt{x+1}$ en torno a $x_0 = 0$. Aproxime $\sqrt{0.5}$, $\sqrt{0.75}$, $\sqrt{1.25}$ y $\sqrt{1.5}$ usando $P_3(x)$, y calcule los errores reales.

9. Determine el segundo polinomio de Taylor $P_2(x)$ para la función $f(x) = e^x \cos x$ en torno a $x_0 = 0$.

a. Use $P_2(0.5)$ para aproximar $f(0.5)$. Determine una cota superior para el error $|f(0.5) - P_2(0.5)|$ por medio de la fórmula para el error y compárelo con el error real.

b. Calcule una cota para el error $|f(x) - P_2(x)|$ al usar $P_2(x)$ para aproximar $f(x)$ en el intervalo $[0, 1]$.

c. Aproxime $\int_0^1 f(x)\,dx$ por medio de $\int_0^1 P_2(x)\,dx$.

d. Calcule una cota superior para el error en (c) mediante $\int_0^1 |R_2(x)\,dx|$ y compárela con el error real.

10. Repita el ejercicio 9 con $x_0 = \pi/6$.

11. Determine el tercer polinomio de Taylor $P_3(x)$ para la función $f(x) = (x-1)\ln x$ respecto a $x_0 = 1$.

a. Use $P_3(0.5)$ para aproximar $f(0.5)$. Determine una cota superior para el error $|f(0.5) - P_3(0.5)|$ por medio de la fórmula para el error y compárelo con el error real.

b. Calcule una cota para el error $|f(x) - P_3(x)|$ al usar $P_3(x)$ para aproximar $f(x)$ en el intervalo $[0.5, 1.5]$.

c. Aproxime $\int_{0.5}^{1.5} f(x)\,dx$ usando $\int_{0.5}^{1.5} P_3(x)\,dx$.

d. Calcule una cota superior para el error en (c) mediante $\int_{0.5}^{1.5} |R_3(x)\,dx|$, y compárela con el error real.

12. Sean $f(x) = 2x\cos(2x) - (x-2)^2$ y $x_0 = 0$.

a. Determine el tercer polinomio de Taylor $P_3(x)$ y úselo para aproximar $f(0.4)$.

b. Use la fórmula del error en el teorema de Taylor y determine con ella una cota superior para el error $|f(0.4) - P_3(0.4)|$. Calcule el error real.

c. Determine el cuarto polinomio de Taylor $P_4(x)$ y úselo para aproximar $f(0.4)$.

d. Con la fórmula del error en el teorema de Taylor determine una cota superior para el error $|f(0.4) - P_4(0.4)|$. Calcule el error real.

13. Calcule el cuarto polinomio de Taylor $P_4(x)$ para la función $f(x) = xe^{x^2}$ en torno a $x_0 = 0$.

a. Calcule una cota superior para $|f(x) - P_4(x)|$, con $0 \le x \le 0.4$.

b. Aproxime $\int_0^{0.4} f(x)\,dx$ usando $\int_0^{0.4} P_4(x)\,dx$.

c. Determine una cota superior para el error en (b) usando $\int_0^{0.4} P_4(x)\,dx$.

d. Aproxime $f'(0.2)$ usando $P_4'(0.2)$ y calcule el error.

14. Use el término del error de un polinomio de Taylor para estimar el error implicado al emplear $\operatorname{sen} x \approx x$ para aproximar sen 1°.

15. Use un polinomio de Taylor respecto a $\pi/4$ para aproximar cos 42° con una precisión de 10^{-6}.

16. Sea $f(x) = e^{x/2}\operatorname{sen}(x/3)$. Utilice Maple para determinar lo siguiente.

a. El tercer polinomio de Maclaurin $P_3(x)$.

b. $f^{(4)}(x)$ y una cota para el error $|f(x) - P_3(x)|$ en $[0, 1]$.

17. Sea $f(x) = \ln(x^2 + 2)$. Utilice Maple para determinar lo siguiente.

a. El polinomio de Taylor $P_3(x)$ para f desarrollada en torno a $x_0 = 1$.

b. El error máximo $|f(x) - P_3(x)|$ para $0 \le x \le 1$.

c. El polinomio de Maclaurin $\tilde{P}_3(x)$ para f.

d. El error máximo $|f(x) - P_3(x)|$ para $0 \le x \le 1$.

e. ¿$P_3(0)$ aproxima a $f(0)$ mejor de lo que $\tilde{P}_3(1)$ aproxima a $f(1)$?

18. Sean $f(x) = (1-x)^{-1}$ y $x_0 = 0$. Determine el n-ésimo polinomio de Taylor $P_n(x)$ para $f(x)$ en torno a x_0. Determine un valor de n necesario para que $P_n(x)$ aproxime a $f(x)$ hasta 10^{-6} en $[0, 0.5]$.

19. Sean $f(x) = e^x$ y $x_0 = 0$. Determine el n-ésimo polinomio de Taylor $P_n(x)$ para $f(x)$ respecto a x_0. Determine un valor de n necesario para que $P_n(x)$ aproxime a $f(x)$ hasta 10^{-6} en $[0, 0.5]$.

20. Obtenga el n-ésimo polinomio de Maclaurin $P_n(x)$ para $f(x) = \arctan x$.

21. El polinomio $P_2(x) = 1 - \frac{1}{2}x^2$ se usará para aproximar $f(x) = \cos x$ en $[-\frac{1}{2}, \frac{1}{2}]$. Determine una cota para el error máximo.

22. El n-ésimo polinomio de Taylor para una función f en x_0 se conoce a veces como un polinomio de grado n a lo sumo, que aproxima "mejor" a f cerca de x_0.

a. Explique por qué es adecuada esta descripción.

b. Determine el polinomio cuadrático que mejor aproxima una función f cerca de $x_0 = 1$ si la recta tangente en $x_0 = 1$ tiene la ecuación $y = 4x - 1$ y si $f''(1) = 6$.

23. Para obtener la aproximación 2.5 de e se utiliza un polinomio de Maclaurin para e^x. La cota del error en esta aproximación se calcula como $E = \frac{1}{6}$. Determine una cota para el error en E.

24. La *función error* definida como

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

proporciona la probabilidad de que cualquiera de una serie de ensayos esté a menos de x unidades de la media, suponiendo que los ensayos tienen una distribución normal con media 0 y desviación estándar $\sqrt{2}/2$. Esta integral no se puede evaluar en términos de funciones elementales, de modo que se debe usar una técnica de aproximación.

a. Integre la serie de Maclaurin para e^{-x^2} con el fin de mostrar que

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)k!}$$

b. La función error también se puede expresar en la forma

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} \sum_{k=0}^{\infty} \frac{2^k x^{2k+1}}{1 \cdot 3 \cdot 5 \cdots (2k+1)}.$$

Verifique que las dos series coinciden para $k = 1, 2, 3$ y 4. [*Sugerencia*: Use la serie de Maclaurin para e^{-x^2}.]

c. Use la serie del inciso (a) para aproximar erf(1) hasta 10^{-7}.

d. Use el mismo número de términos del inciso (c) para aproximar erf(1) con la serie del inciso (b).

e. Explique por qué hay dificultades al usar la serie del inciso (b) para aproximar erf(x).

25. Una función $f: [a, b] \to \mathbb{R}$ satisface una *condición de Lipschitz* con constante de Lipschitz L en $[a, b]$ si, para cada $x, y \in [a, b]$, tenemos $|f(x) - f(y)| \le L|x - y|$.

a. Demuestre que si f satisface una condición de Lipschitz con constante de Lipschitz en un intervalo $[a, b]$, entonces $f \in C[a, b]$.

b. Demuestre que si f tiene una derivada acotada en $[a, b]$ por L, entonces f satisface una condición de Lipschitz con constante de Lipschitz L en $[a, b]$.

c. Dé un ejemplo de una función que es continua en un intervalo cerrado pero que no satisfaga una condición de Lipschitz en el intervalo.

26. Suponga que $f \in C[a, b]$, que x_1 y x_2 están en $[a, b]$ y que c_1 y c_2 son constantes positivas. Demuestre que existe un número ξ entre x_1 y x_2 tal que

$$f(\xi) = \frac{c_1 f(x_1) + c_2 f(x_2)}{c_1 + c_2}.$$

27. Sean $f \in C[a, b]$ y p en el intervalo abierto (a, b).

a. Suponga que $f(p) \neq 0$. Demuestre que existe $\delta > 0$ tal que $f(x) \neq 0$ para toda x en $[p - \delta, p + \delta]$, donde $[p - \delta, p + \delta]$ es subconjunto de $[a, b]$.

b. Suponga que $f(p) = 0$ y sea $k > 0$ un valor dado. Demuestre que existe $\delta > 0$ tal que $|f(x)| \le k$ para toda x en $[p - \delta, p + \delta]$, donde $[p - \delta, p + \delta]$ es subconjunto de $[a, b]$.

1.2 Errores de redondeo y aritmética de una computadora

La aritmética que realiza una calculadora o una computadora es distinta de la aritmética de nuestros cursos de álgebra o cálculo. Por experiencia esperaría que siempre se tuviera como enunciados verdaderos cosas como $2 + 2 = 4$, $4 \cdot 4 = 16$ y $(\sqrt{3})^2 = 3$. En la aritmética computacional común, siempre se tendrán los dos primeros, pero no siempre el tercero. Para ver por qué, debemos explorar el mundo de la aritmética con un número finito de cifras.

En nuestro mundo matemático tradicional permitimos la existencia de números con una cantidad infinita de cifras. La aritmética que usamos en este mundo define a $\sqrt{3}$ como el único número positivo tal que al multiplicarse por sí mismo produce el entero 3. Sin embargo, en el mundo de las computadoras, cada número representable tiene sólo un número finito, fijo, de cifras. Esto significa, por ejemplo, que sólo los números racionales (y no todos ellos) se pueden representar con exactitud. Puesto que $\sqrt{3}$ no es racional, se da una representación aproximada, una cuyo cuadrado no será 3, aunque sí lo bastante cercano a 3 como para que sea aceptable en la mayor parte de las situaciones. Luego, en muchos casos, esta aritmética de la máquina es satisfactoria y se aprueba sin más, aunque a veces esta discrepancia puede generar problemas.

Los *errores de redondeo* surgen al usar una calculadora o computadora para cálculos con números reales, pues la aritmética de la máquina sólo utiliza números con una cantidad finita de cifras, de modo que los cálculos se realizan únicamente con representaciones aproximadas de los números verdaderos. En una computadora común, sólo se usa un subconjunto relativamente pequeño del sistema de números reales para representarlos a todos. Este subconjunto contiene sólo números racionales (tanto positivos como negativos) y almacena la parte fraccionaria, junto con una parte exponencial.

En 1985, el Institute for Electrical and Electronic Engineers, IEEE (Instituto para Ingenieros Eléctricos y Electrónicos) publicó un informe llamado *Binary Floating Point Arithmetic Standard 754-1985*. Se especificaron los formatos para las precisiones simple, doble y extendida; en general, los fabricantes de microcomputadoras utilizan estos estándares para el hardware de punto flotante. Por ejemplo, el coprocesador numérico de las PC utiliza una representación de 64 bits (dígitos binarios) para un número real, llamado *real largo*. El primer bit es un indicador de signo, denotado como s. Le sigue un exponente de 11 bits, c, denominado **característica** y una fracción binaria de 52 bits, f, llamada **mantisa**. La base para el exponente es 2.

Como 52 dígitos binarios corresponden a entre 16 y 17 dígitos decimales, podemos suponer que un número representado en este sistema tiene al menos 16 cifras decimales de precisión. El exponente de 11 dígitos binarios proporciona un intervalo de 0 a $2^{11} - 1 = 2047$. Sin embargo, el uso exclusivo de enteros positivos para el exponente no permitiría una representación adecuada de los números con magnitud pequeña. Para garantizar que estos números también sean representables, se resta 1023 de la característica, de modo que el intervalo del exponente es en realidad de -1023 a 1024.

Para ahorrar espacio de almacenamiento y proporcionar una representación única de cada número de punto flotante, se impone una normalización. El uso de este sistema proporciona un número de punto flotante de la forma

$$(-1)^s 2^{c-1023}(1+f).$$

Considere, por ejemplo, el número de máquina

0 10000000011 1011100100010000000000000000000000000000000000000000.

El bit de la extrema izquierda es cero, lo cual indica que el número es positivo. Los siguientes 11 bits, 10000000011, que dan la característica, son equivalentes al número decimal

$$c = 1 \cdot 2^{10} + 0 \cdot 2^9 + \cdots + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 = 1024 + 2 + 1 = 1027.$$

La parte exponencial del número es, por tanto, $2^{1027-1023} = 2^4$. Los últimos 52 bits especifican que la mantisa es

$$f = 1 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^3 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^5 + 1 \cdot \left(\frac{1}{2}\right)^8 + 1 \cdot \left(\frac{1}{2}\right)^{12}.$$

En consecuencia, este número de máquina representa con precisión al número decimal

$$(-1)^s 2^{c-1023}(1+f) = (-1)^0 \cdot 2^{1027-1023}\left(1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096}\right)\right)$$
$$= 27.56640625.$$

Sin embargo, el siguiente número de máquina menor es

0 10000000011 1011100100001111111111111111111111111111111111111111,

y el siguiente número de máquina mayor es

0 10000000011 1011100100010000000000000000000000000000000000000001.

Esto significa que nuestro número de máquina original representa no sólo a 27.56640625, sino también a la mitad de los números reales que están entre 27.56640625 y los dos números de máquina más próximos a él. Para ser precisos, representa a cualquier número real en el intervalo

[27.5664062499999982236431605997495353221893310546875,
27.5664062500000017763568394002504646778106689453125).

El menor número positivo normalizado que puede representarse tiene ceros en todas partes, excepto en el bit del extremo derecho (donde tiene un 1) y es equivalente a

$$2^{-1023} \cdot (1 + 2^{-52}) \approx 10^{-308},$$

y el mayor tiene un cero al principio seguido de unos; es equivalente a

$$2^{1024} \cdot (2 - 2^{-52}) \approx 10^{308}.$$

Los números que aparecen en los cálculos y tienen una magnitud menor que $2^{-1023} \cdot (1+2^{-52})$ producen un **desbordamiento de la capacidad mínima o subdesbordamiento** y, por lo general, se igualan a cero. Los números mayores que $2^{1024} \cdot (2-2^{-52})$ producen un **desbordamiento*** y hacen que se detengan los cálculos.

El uso de dígitos binarios tiende a encubrir las dificultades de cálculo que aparecen al usar una colección finita de números de máquina para representar a todos los números reales. Para examinar estos problemas, supondremos, para mayor claridad, que los números de máquina se representan en la forma de punto flotante *decimal* normalizada

$$\pm 0.d_1 d_2 \ldots d_k \times 10^n, \qquad 1 \leq d_1 \leq 9, \quad \text{y} \quad 0 \leq d_i \leq 9,$$

para cada $i = 1, 2, \ldots, k$. Los números de esta forma se llaman *números de máquina decimales con k dígitos.*

* N. del R.T.: Estos conceptos se conocen también como: *overflow* y *underflow*.

Cualquier número positivo real dentro del intervalo numérico de la máquina se puede normalizar como

$$y = 0.d_1d_2\ldots d_kd_{k+1}d_{k+2}\ldots \times 10^n.$$

La forma de punto flotante de y, que denotamos $fl(y)$, se obtiene terminando la mantisa de y en k cifras decimales. Hay dos formas de realizar esto. Un método, llamado **truncamiento**, consiste simplemente en cortar los dígitos $d_{k+1}d_{k+2}\ldots$ para obtener

$$fl(y) = 0.d_1d_2\ldots d_k \times 10^n.$$

El otro método, llamado **redondeo**, suma $5 \times 10^{n-(k+1)}$ a y y luego trunca el resultado para obtener un número de la forma

$$fl(y) = 0.\delta_1\delta_2\ldots\delta_k \times 10^n.$$

Así, al redondear, si $d_{k+1} \geq 5$, sumamos 1 a d_k para obtener $fl(y)$; es decir, *redondeamos hacia arriba*. Si $d_{k+1} < 5$, simplemente truncamos todo excepto los primeros k dígitos; así, *redondeamos hacia abajo*. Si el redondeo es hacia abajo, entonces $\delta_i = d_i$ para cada $i = 1, 2, \ldots, k$. Sin embargo, los dígitos podrían cambiar si el redondeo es hacia arriba.

EJEMPLO 1 El número π tiene un desarrollo decimal infinito de la forma $\pi = 3.14159265\ldots$ Escrito en forma decimal normalizada, tenemos

$$\pi = 0.314159265\ldots \times 10^1.$$

La forma de punto flotante de π con un truncamiento a cinco cifras es

$$fl(\pi) = 0.31415 \times 10^1 = 3.1415.$$

Puesto que la sexta cifra del desarrollo decimal de π es 9, la forma de punto flotante de π con un redondeo a cinco cifras es

$$fl(\pi) = (0.31415 + 0.00001) \times 10^1 = 3.1416. \quad \blacksquare$$

El error que resulta al sustituir un número por su forma de punto flotante es el **error de redondeo** (sin importar si se determinó por truncamiento o redondeo). En la siguiente definición se describen dos métodos para medir errores de aproximación.

Definición 1.15 Si p^* es una aproximación de p, el **error absoluto** es $|p - p^*|$ y el **error relativo** es $\frac{|p-p^*|}{|p|}$, siempre que $p \neq 0$. $\blacksquare$

Considere los errores absoluto y relativo al representar p por p^* en el ejemplo siguiente .

EJEMPLO 2 **a.** Si $p = 0.3000 \times 10^1$ y $p^* = 0.3100 \times 10^1$, el error absoluto es 0.1 y el error relativo es $0.333\overline{3} \times 10^{-1}$.

b. Si $p = 0.3000 \times 10^{-3}$ y $p^* = 0.3100 \times 10^{-3}$, el error absoluto es 0.1×10^{-4} y el error relativo es $0.333\bar{3} \times 10^{-1}$.

c. Si $p = 0.3000 \times 10^{4}$ y $p^* = 0.3100 \times 10^{4}$, el error absoluto es 0.1×10^{3} y el error relativo es de nuevo $0.333\bar{3} \times 10^{-1}$.

Este ejemplo muestra que ocurre el mismo error relativo, $0.333\bar{3} \times 10^{-1}$, para una gran variedad de errores absolutos. Como una medida de la precisión, el error absoluto puede llevar a confusiones, en tanto que el error relativo es más significativo, pues toma en cuenta el tamaño del valor. ■

La siguiente definición utiliza el error relativo como una medida de las cifras significativas de precisión para una aproximación.

Definición 1.16 El número p^* aproxima a p con t **cifras significativas** si t es el mayor entero no negativo para el cual

$$\frac{|p - p^*|}{|p|} < 5 \times 10^{-t}.$$ ■

En la tabla 1.1 se ilustra la naturaleza continua de las cifras significativas al enumerar, para diversos valores de p, la mínima cota superior de $|p - p^*|$, que se denota máx $|p - p^*|$, cuando p^* coincide con p hasta cuatro cifras significativas.

Tabla 1.1.

p	0.1	0.5	100	1000	5000	9990	10000
máx $\|p - p^*\|$	0.00005	0.00025	0.05	0.5	2.5	4.995	5.

De regreso a la representación de los números en la máquina, vemos que $fl(y)$ como punto flotante para el número y tiene el error relativo

$$\left|\frac{y - fl(y)}{y}\right|.$$

Si se usan k cifras decimales y el truncamiento para la representación en la máquina de

$$y = 0.d_1d_2\ldots d_kd_{k+1}\ldots \times 10^n,$$

entonces

$$\left|\frac{y - fl(y)}{y}\right| = \left|\frac{0.d_1d_2\ldots d_kd_{k+1}\ldots \times 10^n - 0.d_1d_2\ldots d_k \times 10^n}{0.d_1d_2\ldots \times 10^n}\right|$$

$$= \left|\frac{0.d_{k+1}d_{k+2}\ldots \times 10^{n-k}}{0.d_1d_2\ldots \times 10^n}\right| = \left|\frac{0.d_{k+1}\,d_{k+2}\ldots}{0.d_1d_2\ldots}\right| \times 10^{-k}.$$

Como $d_1 \neq 0$, el valor mínimo del denominador es 0.1. El 1 es la cota superior del numerador. En consecuencia,

$$\left|\frac{y - fl(y)}{y}\right| \leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}.$$

De manera similar, una cota para el error relativo al usar la aritmética de redondeo a k cifras es $0.5 \times 10^{-k+1}$. (Véase el ejercicio 24.)

Observe que las cotas para el error relativo al usar la aritmética con k cifras son independientes del número representado. Este resultado se debe a la manera en que se distribuyen los números de máquina a lo largo de la recta real. Debido a la forma exponencial de la característica, se usa la misma cantidad de números decimales de máquina para representar cada uno de los intervalos $[0.1, 1]$, $[1, 10]$ y $[10, 100]$. De hecho, dentro de los límites de la máquina, la cantidad de números decimales de máquina en $[10^n, 10^{n+1}]$ es constante para todo entero n.

Además de la representación imprecisa de los números, la aritmética realizada en una computadora no es exacta. La aritmética implica el manejo de los dígitos binarios mediante diversas operaciones de corrimiento, o lógicas. Como la mecánica real de estas operaciones no es pertinente a esta presentación, diseñaremos nuestro propio enfoque de la aritmética de una computadora. Aunque nuestra aritmética no dará la imagen exacta, bastará para explicar los problemas potenciales. (Para una explicación de la mecánica real, el lector deberá consultar textos de ciencias de la computación más orientados a los aspectos técnicos, como [Ma], *Computer System Architecture*.)

Suponga que se tienen las representaciones de punto flotante $fl(x)$ y $fl(y)$ para los números reales x y y, y que los símbolos $\oplus$, $\ominus$, $\otimes$, $\oslash$ representan las operaciones de suma, resta, multiplicación y división en la máquina, respectivamente. Supondremos que se usa una aritmética con un número finito de cifras dada por

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \otimes y = fl(fl(x) \times fl(y)),$$
$$x \ominus y = fl(fl(x) - fl(y)), \quad x \oslash y = fl(fl(x) \div fl(y)).$$

Esto corresponde a realizar la aritmética exacta con las representaciones de punto flotante de x y y, para luego convertir el resultado exacto en su representación de punto flotante con un número finito de cifras.

La aritmética de redondeo se lleva a cabo fácilmente en un sistema de álgebra por computadora. La instrucción de Maple

```
>Digits:=t;
```

hace que toda la aritmética se redondee a t cifras. Por ejemplo, $fl(fl(x)+fl(y))$ se realiza con aritmética de redondeo a t cifras mediante

```
>evalf(evalf(x)+evalf(y));
```

La ejecución de la aritmética de truncamiento a t cifras es más difícil y requiere una serie de pasos o un procedimiento. En el ejercicio 27 se estudia este problema.

EJEMPLO 3 Suponga que $x = \frac{5}{7}$, $y = \frac{1}{3}$ y que se usa el truncamiento a cinco cifras para los cálculos aritméticos donde intervienen x y y. En la tabla 1.2 se enumeran los valores de estas operaciones de tipo computadora con $fl(x) = 0.71428 \times 10^0$ y $fl(y) = 0.33333 \times 10^0$. ■

Tabla 1.2

Operación	Resultado	Valor real	Error absoluto	Error relativo
$x \oplus y$	0.10476×10^1	22/21	0.190×10^{-4}	0.182×10^{-4}
$x \ominus y$	0.38095×10^0	8/21	0.238×10^{-5}	0.625×10^{-5}
$x \otimes y$	0.23809×10^0	5/21	0.524×10^{-5}	0.220×10^{-4}
$y \oslash x$	0.21428×10^1	15/7	0.571×10^{-4}	0.267×10^{-4}

Como el máximo error relativo para las operaciones del ejemplo 3 es 0.267×10^{-4}, la aritmética produce resultados satisfactorios con cinco cifras. Suponga, sin embargo, que también tenemos los números $u = 0.714251$, $v = 98765.9$ y $w = 0.111111 \times 10^{-4}$, de modo que $fl(u) = 0.71425 \times 10^0$, $fl(v) = 0.98765 \times 10^5$ y $fl(w) = 0.11111 \times 10^{-4}$. (Elegimos estos números para ilustrar algunos problemas que pueden surgir con la aritmética cuando se tiene una cantidad finita de cifras.)

En la tabla 1.3, $x \ominus u$ produce un error absoluto pequeño, pero un error relativo grande. La división posterior entre el número pequeño w o la multiplicación por el número grande v aumenta el error absoluto, sin modificar el error relativo. La suma de los números grande y pequeño u y v produce un error absoluto grande pero no uno relativo similar.

Tabla 1.3

Operación	Resultado	Valor real	Error absoluto	Error relativo
$x \ominus u$	0.30000×10^{-4}	0.34714×10^{-4}	0.471×10^{-5}	0.136
$(x \ominus u) \oslash w$	0.29629×10^1	0.34285×10^1	0.465	0.136
$(x \ominus u) \otimes v$	0.29629×10^1	0.34285×10^1	0.465	0.136
$u \oplus v$	0.98765×10^5	0.98766×10^5	0.161×10^1	0.163×10^{-4}

Uno de los los cálculos más comunes que producen errores tiene que ver con la cancelación de cifras significativas debido a la resta de números casi iguales. Suponga que dos números casi iguales x y y, con $x > y$, tienen las representaciones de k cifras

$$fl(x) = 0.d_1d_2 \dots d_p\alpha_{p+1}\alpha_{p+2} \dots \alpha_k \times 10^n,$$

y

$$fl(y) = 0.d_1d_2 \dots d_p\beta_{p+1}\beta_{p+2} \dots \beta_k \times 10^n,$$

La forma de punto flotante de $x - y$ es

$$fl(fl(x) - fl(y)) = 0.\sigma_{p+1}\sigma_{p+2} \dots \sigma_k \times 10^{n-p},$$

donde

$$0.\sigma_{p+1}\sigma_{p+2} \dots \sigma_k = 0.\alpha_{p+1}\alpha_{p+2} \dots \alpha_k - 0.\beta_{p+1}\beta_{p+2} \dots \beta_k.$$

El número de punto flotante utilizado para representar $x - y$ tiene a lo sumo $k - p$ cifras significativas. Sin embargo, en la mayor parte de los dispositivos de cálculo a $x - y$ se le asignarán k cifras, de modo que las últimas p se anularán o serán asignadas al azar. En todos los cálculos posteriores con $x - y$ se tendrá el problema de contar con $k - p$ cifras significativas, pues una cadena de cálculos no es más precisa que su parte más débil.

Si una representación con un número finito de cifras o un cálculo introduce un error, éste aumenta al dividir entre un número con magnitud pequeña (o, en forma equivalente, al multiplicar por un número con magnitud grande). Suponga, por ejemplo, que el número z tiene la aproximación con un número finito de cifras $z + \delta$, donde el error δ surge por la representación o un cálculo anterior. Suponga ahora que dividimos entre $\varepsilon = 10^{-n}$, donde $n > 0$. Entonces

$$\frac{z}{\varepsilon} \approx fl\left(\frac{fl(z)}{fl(\varepsilon)}\right) = (z + \delta) \times 10^n.$$

Así, el error absoluto en esta aproximación, $|\delta| \times 10^n$, es el error absoluto original, $|\delta|$, multiplicado por el factor 10^n.

EJEMPLO 4 Sean $p = 0.54617$ y $q = 0.54601$. El valor exacto de $r = p - q$ es $r = 0.00016$. Suponga que la resta se realiza con una aritmética de cuatro cifras. Al redondear p y q a cuatro cifras, tenemos $p^* = 0.5462$ y $q^* = 0.5460$, respectivamente, y $r^* = p^* - q^* = 0.0002$ es la aproximación de cuatro cifras de r. Como

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0002|}{|0.00016|} = 0.25,$$

el resultado sólo tiene una cifra significativa, en tanto que la precisión para p^* y q^* fue de cuatro y cinco cifras significativas, respectivamente.

Si se usa el truncamiento para obtener las cuatro cifras, las aproximaciones de cuatro cifras de p, q y r son $p^* = 0.5461$, $q^* = 0.5460$ y $r^* = p^* - q^* = 0.0001$. Con esto se obtiene

$$\frac{|r - r^*|}{|r|} = \frac{|0.00016 - 0.0001|}{|0.00016|} = 0.375,$$

lo que también produce sólo una cifra de precisión. ■

La pérdida de precisión debida al error de redondeo se puede evitar a menudo mediante la reformulación del problema, como se muestra en el siguiente ejemplo.

EJEMPLO 5 La fórmula cuadrática establece que las raíces de $ax^2 + bx + c = 0$, cuando $a \neq 0$, son

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{y} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{1.1}$$

Con aritmética de redondeo a cuatro cifras, considere esta fórmula aplicada a la ecuación $x^2 + 62.10x + 1 = 0$, cuyas raíces son aproximadamente

$$x_1 = -0.01610723 \quad \text{y} \quad x_2 = -62.08390.$$

En esta ecuación, b^2 es mucho mayor que $4ac$, de modo que el numerador en el cálculo de x_1 implica la *resta* de números casi iguales. Como

$$\sqrt{b^2 - 4ac} = \sqrt{(62.10)^2 - (4.000)(1.000)(1.000)} = \sqrt{3856. - 4.000} = \sqrt{3852.}$$
$$= 62.06,$$

tenemos

$$fl(x_1) = \frac{-62.10 + 62.06}{2.000} = \frac{-0.04000}{2.000} = -0.02000,$$

una mala aproximación a $x_1 = -0.01611$, con el error relativo grande

$$\frac{|-0.01611 + 0.02000|}{|-0.01611|} \approx 2.4 \times 10^{-1}.$$

Por otro lado, el cálculo de x_2 implica la *suma* de los números casi iguales $-b$ y $-\sqrt{b^2 - 4ac}$. Esto no presenta problemas, pues

$$fl(x_2) = \frac{-62.10 - 62.06}{2.000} = \frac{-124.2}{2.000} = -62.10$$

tiene el error relativo pequeño

$$\frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \times 10^{-4}.$$

Para obtener una aproximación más precisa con redondeo a cuatro cifras para x_1, se cambia la forma de la fórmula cuadrática mediante la *racionalización del numerador*:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right) = \frac{b^2 - (b^2 - 4ac)}{2a(-b - \sqrt{b^2 - 4ac})},$$

lo que se simplifica como una fórmula cuadrática alternativa

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}. \tag{1.2}$$

Al usar (1.2) tenemos

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.01610,$$

con el pequeño error relativo 6.2×10^{-4}. ■

La técnica de racionalización se puede aplicar también para obtener la siguiente fórmula cuadrática alternativa para x_2:

$$x_2 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}. \tag{1.3}$$

Esta forma se utilizará si b es un número negativo. Sin embargo, en el ejemplo 5, el uso incorrecto de esta fórmula para x_2 no sólo produciría la resta de números casi iguales, sino también la división entre el resultado pequeño de esta resta. La imprecisión que esta combinación produce,

$$fl(x_2) = \frac{-2c}{b - \sqrt{b^2 - 4ac}} = \frac{-2.000}{62.10 - 62.06} = \frac{-2.000}{0.04000} = -50.00,$$

tiene el gran error relativo 1.9×10^{-1}.

La pérdida de precisión debida a un error de redondeo también se puede reducir reordenando los cálculos, como se muestra en el siguiente ejemplo.

EJEMPLO 6 Evalúe $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ en $x = 4.71$ con una aritmética de tres cifras.

En la tabla 1.4 se dan los resultados intermedios de los cálculos. Verifique con cuidado estos resultados para asegurarse de que es correcto su concepto de aritmética con un número finito de cifras. Observe que los valores truncados a tres cifras sólo conservan las tres primeras, sin ningún redondeo; estos valores difieren de manera significativa de los valores redondeados a tres cifras.

Tabla 1.4

	x	x^2	x^3	$6.1x^2$	$3.2x$
Exacto	4.71	22.1841	104.487111	135.32301	15.072
Tres cifras (truncamiento)	4.71	22.1	104.	134.	15.0
Tres cifras (redondeo)	4.71	22.2	105.	135.	15.1

$$\begin{aligned}
\text{Exacto:} \quad & f(4.71) = 104.487111 - 135.32301 + 15.072 + 1.5 \\
& \qquad = -14.263899; \\
\text{Tres cifras (truncamiento):} \quad & f(4.71) = ((104. - 134.) + 15.0) + 1.5 = -13.5; \\
\text{Tres cifras (redondeo):} \quad & f(4.71) = ((105. - 135.) + 15.1) + 1.5 = -13.4.
\end{aligned}$$

Los errores relativos para los métodos con tres cifras son

$$\left|\frac{-14.263899 + 13.5}{-14.263899}\right| \approx 0.05, \quad \text{para truncamiento}$$

y

$$\left|\frac{-14.263899 + 13.4}{-14.263899}\right| \approx 0.06, \quad \text{para redondeo.}$$

Como método alternativo, $f(x)$ se puede escribir de una manera **anidada** como

$$f(x) = x^3 - 6.1x^2 + 3.2x + 1.5 = ((x - 6.1)x + 3.2)x + 1.5.$$

Esto da como resultado

$$\text{Tres cifras (truncamiento):} \quad f(4.71) = ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 = -14.2$$

y una respuesta con redondeo a tres cifras de -14.3. Los nuevos errores relativos son

$$\text{Tres cifras (truncamiento):} \quad \left|\frac{-14.263899 + 14.2}{-14.263899}\right| \approx 0.0045;$$

$$\text{Tres cifras (redondeo):} \quad \left|\frac{-14.263899 + 14.3}{-14.263899}\right| \approx 0.0025.$$

El anidamiento redujo el error relativo para la aproximación por truncamiento a menos de 10% del original. Para la aproximación por redondeo, la mejora ha sido más drástica; el error en este caso se redujo más de 95%. ■

Los polinomios *siempre* deben expresarse en forma anidada antes de realizar cualquier evaluación, pues esta forma minimiza el número de cálculos aritméticos. La disminución del error en el ejemplo 6 se debe a la reducción de los cálculos, de cuatro multiplicaciones y tres sumas a dos multiplicaciones y tres sumas. Una forma de reducir los errores de redondeo consiste en reducir el número de cálculos que pueden producir errores.

CONJUNTO DE EJERCICIOS 1.2

1. Calcule el error absoluto y el error relativo en las aproximaciones de p mediante p^*.

   a. $p = \pi, p^* = 22/7$ b. $p = \pi, p^* = 3.1416$

   c. $p = e, p^* = 2.718$ d. $p = \sqrt{2}, p^* = 1.414$

e. $p = e^{10}, p^* = 22000$ **f.** $p = 10^{\pi}, p^* = 1400$

g. $p = 8!, p^* = 39900$ **h.** $p = 9!, p^* = \sqrt{18\pi}\,(9/e)^9$

2. Determine el mayor intervalo en que debe estar p^* para aproximar p con un error relativo de a lo sumo 10^{-4} para cada valor de p.

a. π **b.** e **c.** $\sqrt{2}$ **d.** $\sqrt[3]{7}$

3. Suponga que p^* debe aproximar a p con un error relativo a lo sumo 10^{-3}. Determine el máximo intervalo en que debe estar p^* para cada valor de p.

a. 150 **b.** 900 **c.** 1500 **d.** 90

4. Realice los siguientes cálculos (i) en forma exacta, (ii) mediante una aritmética de truncamiento a tres cifras y (iii) con una aritmética de redondeo a tres cifras. (iv) Calcule los errores relativos en los incisos (ii) y (iii).

a. $\frac{4}{5} + \frac{1}{3}$ **b.** $\frac{4}{5} \cdot \frac{1}{3}$

c. $\left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20}$ **d.** $\left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20}$

5. Use una aritmética de redondeo a tres cifras para los siguientes cálculos. Calcule el error absoluto y el error relativo con el valor exacto determinado a por lo menos cinco cifras.

a. $133 + 0.921$ **b.** $133 - 0.499$

c. $(121 - 0.327) - 119$ **d.** $(121 - 119) - 0.327$

e. $\dfrac{\frac{13}{14} - \frac{6}{7}}{2e - 5.4}$ **f.** $-10\pi + 6e - \frac{3}{62}$

g. $\left(\frac{2}{9}\right) \cdot \left(\frac{9}{7}\right)$ **h.** $\dfrac{\pi - \frac{22}{7}}{\frac{1}{17}}$

6. Repita el ejercicio 5 usando una aritmética de redondeo a cuatro cifras.

7. Repita el ejercicio 5 con una aritmética de truncamiento a tres cifras.

8. Repita el ejercicio 5 mediante una aritmética de truncamiento a cuatro cifras.

9. Los primeros tres términos no nulos de la serie de Maclaurin para la función arco tangente son $x - (1/3)x^3 + (1/5)x^5$. Calcule el error absoluto y el error relativo en las siguientes aproximaciones de π usando el polinomio en vez de la función arco tangente.

a. $4\left[\arctan\left(\frac{1}{2}\right) + \arctan\left(\frac{1}{3}\right)\right]$ **b.** $16 \arctan\left(\frac{1}{5}\right) - 4 \arctan\left(\frac{1}{239}\right)$

10. El número e se puede definir como $e = \sum_{n=0}^{\infty} (1/n!)$, donde $n! = n(n-1)\cdots 2 \cdot 1$ para $n \neq 0$ y $0! = 1$. Calcule el error absoluto y el error relativo en las siguientes aproximaciones de e:

a. $\sum_{n=0}^{5} \frac{1}{n!}$ **b.** $\sum_{n=0}^{10} \frac{1}{n!}$

11. Sea

$$f(x) = \frac{x \cos x - \operatorname{sen} x}{x - \operatorname{sen} x}.$$

a. Calcule $\lim_{x \to 0} f(x)$.

b. Use una aritmética de redondeo a cuatro cifras para evaluar $f(0.1)$.

c. Reemplace cada función trigonométrica por su tercer polinomio de Maclaurin y repita el inciso (b).

d. El valor real es $f(0.1) = -1.99899998$. Determine el error relativo para los valores obtenidos en los incisos (b) y (c).

12. Sea

$$f(x) = \frac{e^x - e^{-x}}{x}.$$

a. Calcule $\lim_{x \to 0} (e^x - e^{-x})/x$.

b. Use una aritmética de redondeo a tres cifras para evaluar $f(0.1)$.

c. Reemplace cada función exponencial por su tercer polinomio de Maclaurin y repita el inciso (b).

d. El valor real es $f(0.1) = 2.003335000$. Determine el error relativo para los valores obtenidos en los incisos (a) y (b).

13. Use una aritmética de redondeo a cuatro cifras y las fórmulas del ejemplo 5 para determinar las aproximaciones más precisas de las raíces de las siguientes ecuaciones cuadráticas. Calcule los errores absolutos y los errores relativos.

a. $\frac{1}{3}x^2 - \frac{123}{4}x + \frac{1}{6} = 0$ **b.** $\frac{1}{3}x^2 + \frac{123}{4}x - \frac{1}{6} = 0$

c. $1.002x^2 - 11.01\,x + 0.01265 = 0$ **d.** $1.002x^2 + 11.01x + 0.01265 = 0$

14. Repita el ejercicio 13 con una aritmética de truncamiento a cuatro cifras.

15. Use el formato real largo de 64 bits para determinar el decimal equivalente a los siguientes números de máquina de punto flotante.

a. 0 10000001010 1001001000000000000000000000000000000000000000000000

b. 1 10000001010 1001001000000000000000000000000000000000000000000000

c. 0 01111111111 0101001000000000000000000000000000000000000000000000

d. 0 01111111111 0101001000000000000000000000000000000000000000000000

16. Determine la forma decimal de los números de máquina más próximos (mayor y menor) a los números dados en el ejercicio 15.

17. Suponga que dos puntos (x_0, y_0) y (x_1, y_1) están en una línea recta, con $y_1 \neq y_0$. Se tienen dos fórmulas para determinar la ordenada al origen de la recta:

$$x = \frac{x_0 y_1 - x_1 y_0}{y_1 - y_0} \quad \text{y} \quad x = x_0 - \frac{(x_1 - x_0)\, y_0}{y_1 - y_0}.$$

a. Demuestre que ambas fórmulas son algebraicamente correctas.

b. Utilice los datos $(x_0, y_0) = (1.31, 3.24)$ y $(x_1, y_1) = (1.93, 4.76)$ y la aritmética de redondeo a tres cifras para calcular la ordenada al origen de ambas formas. ¿Cuál método es mejor y por qué?

18. El polinomio de Taylor de grado n para $f(x) = e^x$ es $\sum_{i=0}^{n} (x^i/i!)$. Use el polinomio de Taylor de grado nueve y la aritmética de truncamiento a tres cifras para determinar una aproximación de e^{-5} mediante cada uno de los métodos siguientes.

a. $e^{-5} \approx \sum_{i=0}^{9} \frac{(-5)^i}{i!} = \sum_{i=0}^{9} \frac{(-1)^i 5^i}{i!}$ **b.** $e^{-5} = \frac{1}{e^5} \approx \frac{1}{\sum_{i=0}^{9} \frac{5^i}{i!}}$.

c. Un valor aproximado de e^{-5} con tres cifras correctas es 6.74×10^{-3}. ¿Cuál fórmula, (a) o (b), da más precisión? ¿Por qué?

19. El sistema lineal dos por dos

$$ax + by = e,$$
$$cx + dy = f,$$

donde a, b, c, d, e, f están dados, se puede resolver en términos de x y y como sigue:

$$\text{establezca } m = \frac{c}{a}, \quad \text{siempre que } a \neq 0;$$
$$d_1 = d - mb;$$
$$f_1 = f - mc;$$
$$y = \frac{f_1}{d_1};$$
$$x = \frac{(e - by)}{a}.$$

Resuelva los siguientes sistemas lineales mediante una aritmética de redondeo a cuatro cifras.

a. $1.130x - 6.990y = 14.20$ **b.** $1.013x - 6.099y = 14.22$

c. $8.110x + 12.20y = -0.1370$ **d.** $-18.11x + 112.2y = -0.1376$

20. Repita el ejercicio 19 con una aritmética de truncamiento a cuatro cifras.

21. **a.** Demuestre que la técnica de anidamiento de polinomios descrita en el ejemplo 6 también sirve para evaluar

$$f(x) = 1.01e^{4x} - 4.62e^{3x} - 3.11e^{2x} + 12.2e^{x} - 1.99.$$

b. Use una aritmética de redondeo a tres cifras, con el supuesto de que $e^{1.53} = 4.62$ y el hecho de que $e^{nx} = (e^x)^n$ para evaluar $f(1.53)$ según lo dado en el inciso (a).

c. Repita el inciso (b) pero primero anide los cálculos.

d. Compare las aproximaciones de los incisos (b) y (c) con el resultado real hasta tres cifras $f(1.53) = -7.61$.

22. Un paralelepípedo rectangular tiene lados de longitudes 3, 4 y 5 cm, medidos al centímetro más cercano. ¿Cuáles son las mejores cotas inferior y superior para el volumen de este paralelepípedo y cuáles para el área de su superficie?

23. Sea $P_n(x)$ el polinomio de Maclaurin de grado n para la función arco tangente. Por medio de Maple con 75 cifras decimales determine el valor de n necesario para aproximar π dentro de 10^{-25} con las fórmulas siguientes.

a. $4\left[P_n\left(\frac{1}{2}\right) + P_n\left(\frac{1}{3}\right)\right]$ **b.** $16P_n\left(\frac{1}{5}\right) - 4P_n\left(\frac{1}{239}\right)$

24. Suponga que $fl(y)$ es una aproximación de y con un redondeo a k cifras. Demuestre que

$$\left|\frac{y - fl(y)}{y}\right| \leq 0.5 \times 10^{-k+1}.$$

[*Sugerencia*: si $d_{k+1} < 5$, entonces $fl(y) = 0.d_1d_2 \ldots d_k \times 10^n$. Si $d_{k+1} \geq 5$, entonces $fl(y) = 0.d_1d_2 \ldots d_k \times 10^n + 10^{n-k}$.]

25. El coeficiente binomial

$$\binom{m}{k} = \frac{m!}{k!\,(m-k)!}$$

describe el número de formas para elegir un subconjunto de k objetos en un conjunto con m elementos.

a. Suponga que los números decimales de máquina tienen la forma

$\pm 0.d_1d_2d_3d_4 \times 10^n$, con $1 \leq d_1 \leq 9$, $0 \leq d_i \leq 9$ si $i = 2, 3, 4$ y $|n| \leq 15$.

¿Cuál es el mayor valor de m para el que el coeficiente binomial $\binom{m}{k}$, pueda calcularse para cualquier k mediante la definición sin causar desbordamiento?

b. Demuestre que $\binom{m}{k}$ también se puede calcular como

$$\binom{m}{k} = \left(\frac{m}{k}\right)\left(\frac{m-1}{k-1}\right)\cdots\left(\frac{m-k+1}{1}\right).$$

c. ¿Cuál es el mayor valor de m para que el coeficiente binomial $\binom{m}{3}$ pueda calcularse mediante la fórmula del inciso (b) sin causar un sobreflujo?

d. Use la ecuación en (b) y una aritmética de truncamiento a cuatro cifras para calcular el número de manos posibles de 5 cartas en una baraja con 52 cartas. Calcule los errores real y relativo.

26. Sea $f \in C[a, b]$ una función cuya derivada existe en (a, b). Suponga que se evaluará a f en x_0 en (a, b), pero en vez de calcular el valor real $f(x_0)$, el valor aproximado $\tilde{f}(x_0)$, es el valor real de f en $x_0 + \epsilon$, es decir, $\tilde{f}(x_0) = f(x_0 + \epsilon)$.

a. Use el teorema del valor medio para estimar el error absoluto $|f(x_0) - \tilde{f}(x_0)|$ y el error relativo $|f(x_0) - \tilde{f}(x_0)| / |f(x_0)|$, suponiendo que $f(x_0) \neq 0$.

b. Si $\epsilon = 5 \times 10^{-6}$ y $x_0 = 1$, calcule las cotas de los errores absoluto y relativo para

i. $f(x) = e^x$ **ii.** $f(x) = \operatorname{sen} x$

c. Repita el inciso (b) con $\epsilon = (5 \times 10^{-6})x_0$ y $x_0 = 10$.

27. El siguiente procedimiento de Maple trunca un número de punto flotante x en t cifras.

```
chop:=proc(x,t);
    if x=0 then 0
    else
        e:=trunc(evalf(log10(abs(x))));
        if e>0 then e:=e+1 fi;
        x2:=evalf(trunc(x*10^{}(t-e))*10^{}(e-t));
    fi
end;
```

Verifique que el procedimiento funciona para los siguientes valores.

a. $x = 124.031, t = 5$ **b.** $x = 124.036, t = 5$

c. $x = -124.031, t = 5$ **d.** $x = -124.036, t = 5$

e. $x = 0.00653, t = 2$ **f.** $x = 0.00656, t = 2$

g. $x = -0.00653, t = 2$ **h.** $x = -0.00656, t = 2$

28. En el ejemplo inicial de este capítulo se describió un experimento físico relativo a la temperatura de un gas bajo presión. En esta aplicación, teníamos que $P = 1.00$ atmósfera, $V = 0.100\ \text{m}^3$, $N = 0.00420$ moles y $R = 0.08206$. Al despejar T en la ley del gas ideal tenemos

$$T = \frac{PV}{NR} = \frac{(1.00)(0.100)}{(0.00420)(0.008206)} = 290.15\ \text{K} = 17\ ^\circ\text{C}.$$

En el laboratorio se determinó que bajo estas condiciones *la temperatura T era de* 15 °C, y que al duplicar la presión y reducir el volumen a la mitad, T era igual a 19 °C. Suponga que los datos están redondeados con una precisión igual al número de cifras dadas y demuestre que las cifras del laboratorio están dentro de las cotas de precisión de la ley del gas ideal.

1.3 Algoritmos y convergencia

En el texto estaremos analizando procedimientos de aproximación, llamados *algoritmos*, relacionados con series de cálculos. Un **algoritmo** es un procedimiento que describe, sin ambigüedades, una serie finita de pasos a realizar en un orden específico. El objeto del algoritmo es poner en práctica un procedimiento para resolver un problema o aproximarse a una solución del problema.

Usaremos un **seudocódigo** para describir los algoritmos. Este seudocódigo especifica la forma de la entrada por proporcionar y la forma de la salida deseada. No con todos los procedimientos numéricos se obtiene una salida satisfactoria para una entrada elegida de manera arbitraria. Como consecuencia, en cada algoritmo se incorpora una técnica para detenerlo, independiente de la técnica numérica, para evitar ciclos infinitos.

En los algoritmos se usan dos símbolos de puntuación:

Un punto (.) indica el fin de un paso,

el punto y coma (;) separa las tareas dentro de un paso.

Las sangrías se usan para indicar que los grupos de enunciados deben considerarse como una sola entidad.

Las técnicas de formación de ciclos en los algoritmos son controladas por un contador, como por ejemplo,

Para $i = 1, 2, \ldots, n$

Establezca $x_i = a_i + i \cdot h$

o por una condición, como

Mientras $i < N$ ejecute Pasos 3–6.

Para permitir una ejecución condicional, usamos las construcciones estándar

Si ... entonces

o

Si ... entonces

otras

construcciones

Los pasos en los algoritmos siguen las reglas de la construcción estructurada de programas. Los hemos organizado de modo que haya pocas dificultades para traducir el seudocódigo a cualquier lenguaje de programación adecuado para las aplicaciones científicas.

A los algoritmos se les añaden comentarios escritos en cursivas y dentro de paréntesis para distinguirlos de los enunciados de los algoritmos.

EJEMPLO 1 Un algoritmo para calcular

$$\sum_{i=1}^{N} x_1 + x_2 + \cdots + x_N,$$

donde N y los números $x_1, x_2, \ldots, x_N$ están dados, se describe como sigue:

ENTRADA $N, x_1, x_2, \ldots, x_n$

SALIDA $SUMA = \sum_{i=1}^{N} x_i$.

Paso 1 Establezca $SUMA = 0$. *(Inicializa el acumulador.)*

Paso 2 Para $i = 1, 2, \ldots, N$ haga
fijar $SUMA = SUMA + x_i$. *(Agrega el siguiente término.)*

Paso 3 SALIDA ($SUMA$);
PARAR.

EJEMPLO 2 El N-ésimo polinomio de Taylor para $f(x) = \ln x$ desarrollado en torno a $x_0 = 1$ es

$$P_N(x) = \sum_{i=1}^{N} \frac{(-1)^{i+1}}{i}(x-1)^i,$$

y el valor de ln 1.5 con ocho cifras decimales es 0.40546511. Suponga que queremos calcular el valor mínimo de N necesario para que

$$|\ln 1.5 - P_N(1.5)| < 10^{-5}$$

sin usar el término del residuo en el polinomio de Taylor. En los cursos de cálculo aprendimos que si $\sum_{n=1}^{\infty} a_n$ es una serie alternante con límite A cuyos términos disminuyen en magnitud, entonces A y la N-ésima suma parcial $A_N = \sum_{n=1}^{N} a_n$ difieren por menos que la magnitud del $(N + 1)$-ésimo término; es decir,

$$|A - A_N| \leq |a_{N+1}|.$$

El siguiente algoritmo usa esta cota.

ENTRADA valor x, tolerancia TOL, número máximo de iteraciones M.
SALIDA grado N del polinomio o un mensaje de error.
Paso 1 Establezca $N = 1$;
$y = x - 1$;
$SUMA = 0$;
$POTENCIA = y$;
$TÉRMINO = y$;
$SIGNO = -1$. *(Se usa para ejecutar la alternancia de signos.)*

Paso 2 Mientras $N \leq M$ realice los Pasos 3-5.

Paso 3 Fijar $SIGNO = -SIGNO$. *(Alterna los signos.)*
$SUMA = SUMA + SIGNO \cdot TÉRMINO$; *(Acumula los términos.)*
$POTENCIA = POTENCIA \cdot y$;
$TÉRMINO = POTENCIA/(N + 1)$. *(Calcula el siguiente término.)*

Paso 4 Si $|TÉRMINO| < TOL$ entonces (*Verifica la precisión.*)
SALIDA (N);
PARAR. (*El procedimiento tuvo éxito.*)

Paso 5 Fijar $N = N + 1$. (*Prepara la siguiente iteración.*)

Paso 6 SALIDA('El método falló'); (*El procedimiento no tuvo éxito.*)
PARAR.

La entrada de nuestro problema es $x = 1.5$, $TOL = 10^{-5}$ y tal vez $M = 15$. Esta elección de M proporciona una cota superior para el número de cálculos que estamos dispuestos a realizar, reconociendo la posibilidad de que falle el algoritmo si se excede esta cota. El hecho de que la salida sea el valor de N o el mensaje de error depende de la precisión del dispositivo utilizado para realizar los cálculos. ■

Nos interesa elegir métodos que produzcan resultados precisos (según las circunstancias) para una amplia variedad de problemas. Uno de los criterios que siempre trataremos de imponer sobre un algoritmo es que los cambios pequeños en los datos iniciales produzcan otros correspondientes en los resultados finales. Un algoritmo que satisfaga esta propiedad es **estable**; en caso contrario, **inestable**. Algunos algoritmos sólo son estables para ciertas elecciones de datos iniciales; a estos se les llama **condicionalmente estables.** Caracterizaremos las propiedades de estabilidad de los algoritmos siempre que sea posible.

Para continuar nuestro análisis del crecimiento de los errores de redondeo y su relación con la estabilidad de un algoritmo, suponga que se introduce un error de magnitud $E_0 > 0$ en cierta etapa de los cálculos y que se denota por E_n la magnitud del error después de n operaciones sucesivas. Los dos casos que surgen con más frecuencia en la práctica se definen como sigue.

Definición 1.17 Suponga que $E_0 > 0$ denota un error inicial y E_n representa la magnitud de un error después de n operaciones sucesivas. Si $E_n = CnE_0$, donde C es una constante independiente de n, entonces se dice que el crecimiento del error es **lineal**. Si $E_n = C^nE_0$ para alguna $C > 1$, entonces el crecimiento del error se denomina **exponencial**. ■

Normalmente es inevitable el crecimiento lineal del error y, cuando C y E_0 son pequeños, por lo general son aceptables los resultados. Por otro lado, hay que evitar el crecimiento exponencial del error, pues el término C^n crece incluso para valores de n relativamente pequeños. Esto conduce a imprecisiones inaceptables, sin importar el tamaño de E_0. En consecuencia, un algoritmo que exhibe un crecimiento lineal del error es estable, pero no así un algoritmo con crecimiento exponencial del error. (Véase la figura 1.13 de la página 34.)

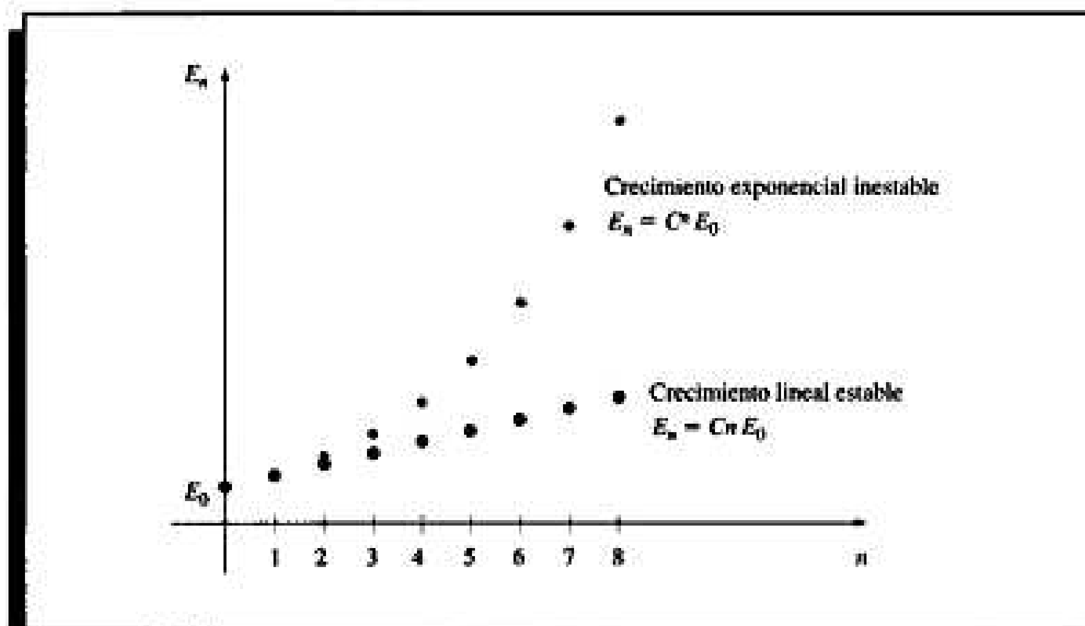
EJEMPLO 3 La ecuación recursiva

$$p_n = \frac{10}{3} p_{n-1} - p_{n-2}, \qquad \text{para } n = 2, 3, \ldots$$

tiene la solución

$$p_n = c_1 \left(\frac{1}{3}\right)^n + c_2 3^n,$$

Figura 1.13



para cualesquiera constantes c_1 y c_2 pues

$$\frac{10}{3}p_{n-1} - p_{n-2} = \frac{10}{3}\left[c_1\left(\frac{1}{3}\right)^{n-1} + c_2 3^{n-1}\right] - \left[c_1\left(\frac{1}{3}\right)^{n-2} + c_2 3^{n-2}\right]$$

$$= c_1\left(\frac{1}{3}\right)^{n-2}\left[\frac{10}{3}\cdot\frac{1}{3} - 1\right] + c_2 3^{n-2}\left[\frac{10}{3}\cdot 3 - 1\right]$$

$$= c_1\left(\frac{1}{3}\right)^{n-2}\left(\frac{1}{9}\right) + c_2 3^{n-2}(9) = c_1\left(\frac{1}{3}\right)^{n} + c_2 3^n = p_n.$$

Si $p_0 = 1$ y $p_1 = \frac{1}{3}$, tenemos $c_1 = 1$ y $c_2 = 0$, de modo que $p_n = (\frac{1}{3})^n$ para toda n. Suponga que usa una aritmética de redondeo a cinco cifras para calcular los términos de la sucesión dada por esta ecuación. Entonces $\hat{p}_0 = 1.0000$ y $\hat{p}_1 = 0.33333$, lo cual requiere modificar las constantes a $\hat{c}_1 = 1.0000$ y $\hat{c}_2 = -0.12500 \times 10^{-5}$. Así que la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ generada está dada por

$$\hat{p}_n = 1.0000\left(\frac{1}{3}\right)^n - 0.12500 \times 10^{-5}(3)^n,$$

y el error de redondeo,

$$p_n - \hat{p}_n = 0.12500 \times 10^{-5}(3^n),$$

crece en forma exponencial con n. Esto se refleja en las imprecisiones extremas después de los primeros términos, como se muestra en la tabla 1.5.

Por otro lado, la ecuación de recurrencia

$$p_n = 2p_{n-1} - p_{n-2}, \quad \text{para } n = 2, 3, \ldots$$

Tabla 1.5

n	$\hat{p}_n$ calculado	p_n correcto	Error relativo
0	0.10000×10^{1}	0.10000×10^{1}	
1	0.33333×10^{0}	0.33333×10^{0}	
2	0.11110×10^{0}	0.11111×10^{0}	9×10^{-5}
3	0.37000×10^{-1}	0.37037×10^{-1}	1×10^{-3}
4	0.12230×10^{-1}	0.12346×10^{-1}	9×10^{-3}
5	0.37660×10^{-2}	0.41152×10^{-2}	8×10^{-2}
6	0.32300×10^{-3}	0.13717×10^{-2}	8×10^{-2}
7	-0.26893×10^{-2}	0.45725×10^{-3}	7×10^{0}
8	-0.92872×10^{-2}	0.15242×10^{-3}	6×10^{1}

tiene la solución $p_n = c_1 + c_2 n$ para las constantes c_1 y c_2 porque

$$2p_{n-1} - p_{n-2} = 2(c_1 + c_2(n-1)) - (c_1 + c_2(n-2))$$
$$= c_1(2-1) + c_2(2n - 2 - n + 2) = c_1 + c_{2n} = p_n.$$

Si $p_0 = 1$ y $p_1 = \frac{1}{3}$, las constantes en esta ecuación se convierten en $c_1 = 1$ y $c_2 = -\frac{2}{3}$, de modo que $p_n = 1 - \frac{2}{3}n$. Con una aritmética de redondeo a cinco cifras se obtiene $\hat{p}_0 = 1.0000$ y $\hat{p}_1 = 0.33333$. En consecuencia, las constantes con redondeo a cinco cifras son $\hat{c}_1 = 1.0000$ y $\hat{c}_2 = -0.66667$. Así,

$$\hat{p}_n = 1.0000 - 0.66667n,$$

y el error por redondeo es

$$p_n - \hat{p}_n = \left(0.66667 - \frac{2}{3}\right)n,$$

que crece linealmente con n. Esto se refleia en la estabilidad que aparece en la tabla 1.6. ■

Tabla 1.6

n	$\hat{p}_n$ calculado	p_n correcto	Error relativo
0	0.10000×10^{1}	0.10000×10^{1}	
1	0.33333×10^{0}	0.33333×10^{0}	
2	-0.33330×10^{0}	-0.33333×10^{0}	9×10^{-5}
3	-0.10000×10^{1}	-0.10000×10^{1}	0
4	-0.16667×10^{1}	-0.16667×10^{1}	0
5	-0.23334×10^{1}	-0.23333×10^{1}	4×10^{-5}
6	-0.30000×10^{1}	-0.30000×10^{1}	0
7	-0.36667×10^{1}	-0.36667×10^{1}	0
8	-0.43334×10^{1}	-0.43333×10^{1}	2×10^{-5}

Los efectos del error por redondeo se pueden reducir mediante una aritmética de orden superior, como la opción de precisión doble o múltiple en la mayor parte de las computadoras. El uso de la aritmética de doble precisión presenta las desventajas de más tiempo de cómputo y no se elimina el crecimiento del error por redondeo, sino solamente se pospone hasta realizar otros cálculos.

Un método para estimar el error de redondeo consiste en usar aritmética de intervalos (es decir, conservar los valores máximo y mínimo en cada paso), de modo que, al final, obtenemos un intervalo que contiene al valor real. Desafortunadamente, se necesitaría un intervalo muy pequeño para una ejecución razonable.

Puesto que se usan con frecuencia las técnicas iterativas relacionadas con sucesiones, esta sección concluye con un breve análisis de cierta terminología usada para describir la rapidez a la que ocurre la convergencia cuando se emplea una técnica numérica. En general, quisiéramos que la técnica convergiese lo más rápido posible. Se usa la siguiente definición para comparar las razones de convergencia de varios métodos.

Definición 1.18 Suponga que $\{\beta_n\}_{n=1}^{\infty}$ es una sucesión cuyo valor de convergencia es cero y que $\{\alpha_n\}_{n=1}^{\infty}$ converge a un número α. Si existe una constante positiva K tal que

$$|\alpha_n - \alpha| \leq K|\beta_n|, \qquad \text{para } n \text{ grande,}$$

entonces decimos que $\{\alpha_n\}_{n=1}^{\infty}$ converge a α con **rapidez de convergencia** $O(\beta_n)$. Esta expresión se lee "O mayúscula de β_n". Se indica escribiendo $\alpha_n = \alpha + O(\beta_n)$. ■

Aunque la definición 1.18 permite comparar $\{\alpha_n\}_{n=1}^{\infty}$ con una sucesión arbitraria $\{\beta_n\}_{n=1}^{\infty}$, en casi todas las situaciones usamos

$$\beta_n = \frac{1}{n^p},$$

para algún número $p > 0$. Por lo general, se tiene interés en el mayor valor de p tal que $\alpha_n = \alpha + O(1/n^p)$.

EJEMPLO 4 Suponga que, para $n \geq 1$,

$$\alpha_n = \frac{n+1}{n^2} \quad \text{y} \quad \hat{\alpha}_n = \frac{n+3}{n^3}.$$

Aunque $\lim_{n\to\infty} \alpha_n = 0$ y $\lim_{n\to\infty} \hat{\alpha}_n = 0$, la sucesión $\{\hat{\alpha}_n\}$ converge en este límite mucho más rápido que la sucesión $\{\alpha_n\}$, usando la aritmética de redondeo a cinco cifras, como se muestra en la tabla 1.7.

Tabla 1.7

n	1	2	3	4	5	6	7
α_n	2.00000	0.75000	0.44444	0.31250	0.24000	0.19444	0.16327
$\hat{\alpha}_n$	4.00000	0.62500	0.22222	0.10938	0.064000	0.041667	0.029155

Si $\beta_n = 1/n$ y $\hat{\beta}_n = 1/n^2$, vemos que

$$|\alpha_n - 0| = \frac{n+1}{n^2} \leq \frac{n+n}{n^2} = 2 \cdot \frac{1}{n} = 2\beta_n$$

y

$$|\hat{\alpha}_n - 0| = \frac{n+3}{n^3} \leq \frac{n+3n}{n^3} = 4 \cdot \frac{1}{n^2} = 4\hat{\beta}_n,$$

de modo que

$$\alpha_n = 0 + O\left(\frac{1}{n}\right) \quad \text{y} \quad \hat{\alpha}_n = 0 + O\left(\frac{1}{n^2}\right).$$

La rapidez de convergencia de $\{\alpha_n\}$ a cero es similar a la convergencia de $\{1/n\}$ a cero, en tanto que $\{\hat{\alpha}_n\}$ converge a cero con una rapidez similar a la de la sucesión $\{1/n^2\}$, la cual converge más rápido. ■

También usamos la notación "O mayúscula" para describir la rapidez de convergencia de funciones.

Definición 1.19 Suponga que $\lim_{h\to 0} G(h) = 0$ y $\lim_{h\to 0} F(h) = L$. Si existe una constante positiva K tal que

$$|F(h) - L| \le K|G(h)|, \qquad \text{para } h \text{ suficientemente pequeña,}$$

entonces escribimos $F(h) = L + O(G(h))$. ■

Por lo común, las funciones que usamos para la comparación tienen la forma $G(h) = h^p$, donde $p > 0$. Nos interesa el mayor valor de p para el que $F(h) = L + O(h^p)$.

EJEMPLO 5 En el ejemplo 3(b) de la sección 1.1 vimos que el tercer polinomio de Taylor da

$$\cos h = 1 - \frac{1}{2}h^2 + \frac{1}{24}h^4 \cos \tilde{\xi}(h),$$

para algún número $\tilde{\xi}(h)$ entre cero y h. En consecuencia,

$$\cos h + \frac{1}{2}h^2 = \frac{1}{24}h^4 \cos \tilde{\xi}(h).$$

De este resultado se deduce que

$$\cos h + \frac{1}{2}h^2 = 1 + O(h^4),$$

pues $|(\cos h + \frac{1}{2}h^2) - 1| = |\frac{1}{24}\cos \tilde{\xi}(h)|h^4 \le \frac{1}{24}h^4$. Esto implica que, cuando $h \to 0$, $\cos h + \frac{1}{2}h^2$ converge a su límite, 1, casi tan rápido como h^4 converge a 0. ■

CONJUNTO DE EJERCICIOS 1.3

1. a. Con aritmética de truncamiento de tres cifras calcule la suma $\sum_{i=1}^{10}(1/i^2)$. Utilice primero $\frac{1}{1} + \frac{1}{4} + \cdots + \frac{1}{100}$ y luego $\frac{1}{100} + \frac{1}{81} + \cdots + \frac{1}{1}$. ¿Cuál método es más preciso? ¿Por qué?

   b. Escriba un algoritmo para sumar la serie finita $\sum_{i=1}^{N} x_i$ en orden inverso.

2. El número e se define como $e = \sum_{n=0}^{\infty}(1/n!)$ donde $n! = n(n-1)\cdots 2\cdot 1$ para $n \neq 0$ y $0! = 1$. Use la aritmética de truncamiento de cuatro cifras para calcular la siguiente aproximación de e y determine los errores absoluto y relativo.

a. $e \approx \sum_{n=0}^{5} \frac{1}{n!}$

b. $e \approx \sum_{j=0}^{5} \frac{1}{(5-j)!}$

c. $e \approx \sum_{n=0}^{10} \frac{1}{n!}$

d. $e \approx \sum_{j=0}^{10} \frac{1}{(10-j)!}$

3. La serie de Maclaurin para la función arctan converge en $-1 < x \leq 1$ y está dada por

$$\arctan x = \lim_{n\to\infty} P_n(x) = \lim_{n\to\infty} \sum_{i=1}^{n} (-1)^{i+1} \frac{x^{2i-1}}{2i-1}.$$

a. Use el hecho de que $\tan \pi/4 = 1$ para determinar el número de términos n de la serie que debemos sumar para garantizar que $|4P_n(1) - \pi| < 10^{-3}$.

b. El lenguaje de programación C++ requiere que el valor aproximado de π esté dentro de 10^{-10}. ¿Cuántos términos de la serie debemos sumar para obtener este grado de precisión?

4. En el ejercicio 3 se detalla un método poco efectivo para obtener una aproximación de π. El método se puede mejorar de manera sustancial si se considera que $\pi/4 = \arctan \frac{1}{2} + \arctan \frac{1}{3}$ y se evalúa la serie para el arctan en $\frac{1}{2}$ y en $\frac{1}{3}$. Calcule el número de términos que debemos sumar para garantizar que la aproximación de π esté dentro de 10^{-3}.

5. Otra fórmula para calcular π se puede deducir de la identidad $\pi/4 = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239}$. Calcule el número de términos que debemos sumar para garantizar que la aproximación de π esté dentro de 10^{-3}.

6. Determine la rapidez de convergencia de las siguientes sucesiones cuando $n \to \infty$.

a. $\lim_{n\to\infty} \operatorname{sen} \frac{1}{n} = 0$

b. $\lim_{n\to\infty} \operatorname{sen} \frac{1}{n^2} = 0$

c. $\lim_{n\to\infty} \left(\operatorname{sen} \frac{1}{n}\right)^2 = 0$

d. $\lim_{n\to\infty} [\ln(n+1) - \ln(n)] = 0$

7. Determine las razones de convergencia de las siguientes funciones cuando $h \to 0$.

a. $\lim_{h\to 0} \frac{\operatorname{sen} h}{h} = 1$

b. $\lim_{h\to 0} \frac{1 - \cos h}{h} = 0$

c. $\lim_{h\to 0} \frac{\operatorname{sen} h - h \cos h}{h} = 0$

d. $\lim_{h\to 0} \frac{1 - e^h}{h} = 1$

8. a. ¿Cuántas multiplicaciones y sumas se necesitan para determinar una suma de la siguiente forma?

$$\sum_{i=1}^{n} \sum_{j=1}^{i} a_i b_j$$

b. Modifique la suma del inciso (a) en una forma equivalente que reduzca el número de cálculos.

9. Suponga que se tiene el polinomio $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ y x_0. Construya un algoritmo para evaluar $P(x_0)$ por medio de una multiplicación anidada.

10. En el ejemplo 5 de la sección 1.2 se dan fórmulas alternativas para las raíces x_1 y x_2 de $ax^2 + bx + c = 0$. Construya un algoritmo con entrada a, b, c y salida x_1, x_2 que permita calcular las raíces x_1 y x_2 (que pueden ser iguales o complejos conjugados) mediante la mejor fórmula para cada raíz.

11. Construya un algoritmo cuya entrada sea un entero $n \geq 1$, los números $x_0, x_1, \ldots, x_n$ y un número x cuya salida sea el producto $(x - x_0)(x - x_1) \cdots (x - x_n)$.

12. Suponga que

$$\frac{1-2x}{1-x+x^2} + \frac{2x-4x^3}{1-x^2+x^4} + \frac{4x^3-8x^7}{1-x^4+x^8} + \cdots = \frac{1+2x}{1+x+x^2},$$

para $x < 1$ y sea $x = 0.25$. Escriba y ejecute un algoritmo que calcule el número de términos necesarios en el miembro izquierdo de la ecuación, de modo que este lado izquierdo difiera del lado derecho en menos de 10^{-6}.

13. **a.** Suponga que $0 < q < p$ y que $\alpha_n = \alpha + O(n^{-p})$. Demuestre que $\alpha_n = \alpha + O(n^{-q})$.

b. Construya una tabla con los valores de $1/n$, $1/n^2$, $1/n^3$ y $1/n^4$ para $n = 5$, 10, 100 y 1000 y analice las variaciones en las tasas de convergencia de estas sucesiones a medida que crece n.

14. **a.** Suponga que $0 < q < p$ y que $F(h) = L + O(h^p)$. Demuestre que $F(h) = L + O(h^q)$.

b. Elabore una tabla con los valores de h, h^2, h^3 y h^4 para $h = 0.5$, 0.1, 0.01 y 0.001, y analice las variaciones en las tasas de convergencia de estas potencias de h cuando h tiende a cero.

15. Suponga que cuando x tiende a cero,

$$F_1(x) = L_1 + O(x^\alpha) \quad \text{y} \quad F_2(x) = L_2 + O(x^\beta).$$

Sean c_1 y c_2 constantes no nulas, y defina

$$F(x) = c_1F_1(x) + c_2F_2(x) \quad \text{y} \quad G(x) = F_1(c_1x) + F_2(c_2x).$$

Demuestre que si $\gamma = \min\{\alpha, \beta\}$, entonces, cuando x tiende a cero,

a. $F(x) = c_1L_1 + c_2L_2 + O(x^\gamma)$ **b.** $G(x) = L_1 + L_2 + O(x^\gamma)$.

16. La sucesión $\{F_n\}$ descrita por $F_0 = 1$, $F_1 = 1$ y $F_{n+2} = F_n + F_{n+1}$, si $n \geq 0$, se llama *sucesión de Fibonacci*. Sus términos aparecen de manera natural en muchas especies botánicas, en particular aquellas cuyos pétalos o escamas están ordenados en la forma de una espiral logarítmica. Considere la sucesión $\{x_n\}$, donde $x_n = F_{n+1}/F_n$. Suponiendo que existe $\lim_{n\to\infty} x_n = x$, demuestre que $x = (1 + \sqrt{5})/2$. Este número se llama la *razón áurea*.

17. La sucesión de Fibonacci también satisface la ecuación

$$F_n \equiv \tilde{F}_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

a. Escriba un procedimiento en Maple para calcular F_{100}.

b. Use Maple con el valor predeterminado por omisión de `Digits` seguido de `evalf` para calcular $\tilde{F}_{100}$.

c. ¿Por qué el resultado del inciso (a) es más preciso que el del inciso (b)?

d. ¿Por qué el resultado del inciso (b) se obtiene más rápido que el del inciso (a)?

e. ¿Qué se obtiene al usar la instrucción `simplify` en vez de `evalf` para calcular $\tilde{F}_{100}$?

18. La serie armónica $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$ diverge, pero la sucesión $\gamma_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n$ converge, pues $\{\gamma_n\}$ es una sucesión acotada no creciente. El límite $\gamma = 0.5772156649\ldots$ de la sucesión $\{\gamma_n\}$ se llama constante de Euler.

a. Use el valor predeterminado de `Digits` en Maple para calcular el valor de n para que γ_n esté a menos de 10^{-2} de γ.

b. Use el valor predeterminado de `Digits` en Maple para calcular el valor de n para que γ_n esté a menos de 10^{-3} de γ.

c. ¿Qué ocurre si se usa el valor predeterminado de `Digits` en Maple para calcular el valor de n para que γ_n esté a menos de 10^{-4} de γ?

1.4 Software numérico

Los paquetes de computación utilizados para aproximar las soluciones numéricas están disponibles en diversas formas. En este libro incluimos los programas escritos en C, FORTRAN, Maple, Mathematica, MATLAB y Pascal con los cuales pueden solucionarse los problemas de los ejemplos y ejercicios. Estos programas darán resultados satisfactorios para la mayor parte de los problemas que necesite resolver, aunque sean de los llamados programas de *propósito específico*. Usamos este término para distinguirlos de los disponibles en las bibliotecas de subrutinas matemáticas. Los programas en estos paquetes se llamarán de *propósito general*.

Distinta, de los algoritmos y programas que se incluyen en este libro, es la finalidad de los programas de propósito general, éstos ofrecen medios para reducir los errores atribuidos al redondeo de la computadora y al desbordamiento tanto de la capacidad mínima como máxima. También describen el intervalo de entradas que conducirá a resultados con cierto nivel de exactitud. Como éstas son características que dependen de la máquina, los paquetes de propósito general usan parámetros que describen las características de punto flotante de la computadora utilizada para realizar los cálculos.

Para ilustrar algunas diferencias entre los programas incluidos en un paquete de propósito general y los programas que empleamos en este libro, consideremos un algoritmo con el que se calcula la norma euclidiana de un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ de dimensión n . Con frecuencia, esta norma se usa en programas de gran tamaño y se define como

$$||\mathbf{x}||_2 = \left[\sum_{i=1}^{n} x_i^2\right]^{1/2}.$$

La norma da una medida para la distancia del vector $\mathbf{x}$ al vector $\mathbf{0}$. Por ejemplo, el vector $\mathbf{x} = (2, 1, 3, -2, -1)^t$ tiene

$$||\mathbf{x}||_2 = [2^2 + 1^2 + 3^2 + (-2)^2 + (-1)^2]^{1/2} = \sqrt{19},$$

de modo que su distancia desde $\mathbf{0} = (0, 0, 0, 0, 0)^t$ es $\sqrt{19} \approx 4.36$.

Veamos un algoritmo del tipo que presentaríamos para este problema. El algoritmo no tiene parámetros que dependan de la máquina ni garantías de precisión, pero dará resultados precisos "la mayor parte de los casos".

ENTRADA $n, x_1, x_2, \ldots, x_n$.

SALIDA *NORMA*.

Paso 1 Tome $SUMA = 0$.

Paso 2 Para $i = 1, 2, \ldots, n$ establezca $SUMA = SUMA + x_i^2$.

Paso 3 Tome $NORMA = SUMA^{1/2}$.

Paso 4 SALIDA (*NORMA*);
PARAR.

Es fácil escribir y comprender un programa basado en este algoritmo. Sin embargo, el programa podría no dar la precisión suficiente por varias razones. Por ejemplo, la magnitud de algunos números podría ser demasiado grande o pequeña para poder representarse con precisión en el sistema de punto flotante de la computadora. Además, el orden usual de los cálculos podría no dar los mejores resultados, o la rutina para obtener la raíz cuadrada del software común podría no ser la más adecuada para el problema. Las cuestiones de este tipo son analizadas por los diseñadores de algoritmos al escribir programas para software de propósito general. Con frecuencia, estos programas se utilizan como subprogramas para resolver problemas mayores, de modo que deben incorporar controles que nosotros no necesitaremos.

Consideremos ahora un algoritmo para un programa de software de propósito general que permita calcular la norma euclidiana. En primer lugar, es posible que un componente x_i del vector esté dentro del rango de la máquina, pero no así el cuadrado del componente. Esto puede ocurrir cuando algún $|x_i|$ es tan pequeño que x_i^2 cause un subdesbordamiento o cuando algún $|x_i|$ es tan grande que x_i^2 cause un desbordamiento. También es posible que todos los términos estén dentro del rango de la máquina, pero que haya un desbordamiento por la suma del cuadrado de uno de los términos a la suma recién calculada.

Como los criterios de exactitud dependen de la computadora utilizada para realizar los cálculos, el algoritmo incorpora algunos parámetros dependientes de la máquina. Suponga que estamos trabajando en una computadora hipotética con base 10, con $t \geq 4$ cifras de precisión, un exponente mínimo *emín*, y un exponente máximo *emáx*. Entonces el conjunto de números de punto flotante en esta máquina consta de 0 y de los números de la forma

$$x = f \cdot 10^e, \qquad \text{donde} \quad f = \pm(f_1 10^{-1} + f_2 10^{-2} + \cdots + f_t 10^{-t}),$$

donde $1 \leq f_1 \leq 9$ y $0 \leq f_i \leq 9$ para cada $i = 2, \ldots, t$, y donde $emín \leq e \leq emáx$. Estas restricciones implican que el número positivo más pequeño representado en esta máquina es $\sigma = 10^{emín-1}$, de modo que cualquier número calculado x con $|x| < \sigma$ provoca un subdesbordamiento y hace que x se iguale a cero. El mayor número positivo es $\lambda = (1 - 10^{-t})10^{emáx}$, y cualquier número calculado x con $|x| > \lambda$ produce un desbordamiento. Cuando hay subdesbordamiento, el programa continúa sin que haya una pérdida significativa de exactitud. Si ocurre un desbordamiento, fallará el programa .

El algoritmo supone que las características de punto flotante de la máquina se describen mediante los parámetros N, s, S, y, Y. N es el número máximo de datos que se pueden sumar con al menos $t/2$ cifras de precisión. Esto significa que el algoritmo tratará de calcular la norma de un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ sólo si $n \leq N$. Para resolver el problema de subdesbordamiento-desbordamiento, los números no nulos de punto flotante se separan en tres grupos: los números x de pequeña magnitud, que satisfacen $0 < |x| < y$; los números x de magnitud media, donde $y \leq |x| < Y$; y los números x de gran magnitud, donde $Y \leq |x|$. Los parámetros y y Y se eligen de modo que no se presenten problemas de subdesbordamiento-desbordamiento al sumar los números de magnitud mediana y elevarlos al cuadrado. Al elevar al cuadrado los números de pequeña magnitud puede ocurrir un subdesbordamiento, por lo cual se usa un factor escalar S mucho mayor que 1, de ese modo $(Sx)^2$ evita el subdesbordamiento aunque no lo haga x^2. Al sumar y elevar al cuadrado los números de gran magnitud se puede tener un desbordamiento, de modo que en este caso se usa un factor de escala mucho menor que 1 para garantizar que $(sx)^2$ no provoque un desbordamiento al calcularlo e incorporarlo en una suma, aun cuando x^2 genere problemas.

Para evitar los cambios de escala innecesarios, y y Y se eligen de modo que el intervalo de números de magnitud media sea lo más amplio posible. El siguiente algoritmo es

una modificación del descrito en [Brow, K, p. 471]. Éste incorpora un procedimiento para cambiar la escala de las componentes del vector que tienen magnitud pequeña hasta encontrar una componente de magnitud mediana. Luego, se invierte la escala de la suma anterior y se continúa elevando al cuadrado y sumando números pequeños y medianos hasta encontrar una componente de gran magnitud. Una vez que aparece ésta, el algoritmo escala la suma anterior y se procede a cambiar la escala, elevar al cuadrado y sumar los números restantes. Supone que al pasar de los números pequeños a los medianos, los números pequeños no escalados resultan insignificantes en comparación con los números medianos. De manera similar, al pasar de números medianos a grandes, los números medianos no escalados son insignificantes comparados con los números grandes. Así, hay que elegir los parámetros de escalamiento de modo que los números sean anulados (igualados a cero) sólo cuando realmente sean insignificantes. Las relaciones usuales entre las características de la máquina descritas por t, σ, λ, *emín*, *emáx* y los parámetros N, s, S, y y Y aparecen después del algoritmo.

En el algoritmo se utilizan tres banderas para indicar las diversas etapas del proceso de suma. Estas banderas reciben sus valores iniciales en el paso 3. La BANDERA 1 es 1 hasta que aparezca un componente mediano o grande; entonces se transforma en 0. La BANDERA 2 es 0 mientras se sumen números pequeños, se convierte en 1 al encontrar un número mediano por primera vez y vuelve a ser 0 al aparecer un número grande. La BANDERA 3 es igual a 0 al principio y se transforma en 1 cuando aparece un número grande por vez primera. En el paso 3 también se introduce la bandera HECHO, que es igual a 0 hasta concluir los cálculos, y, una vez terminados, vuelve a ser 1.

ENTRADA $N, s, S, y, Y, \lambda, n, x_1, x_2, \ldots, x_n$.

SALIDA *NORMA* o un mensaje pertinente de error.

Paso 1 Si $n \leq 0$ entonces SALIDA ('El entero n debe ser positivo'.);
PARAR.

Paso 2 Si $n \geq N$ entonces SALIDA ('El entero n es demasiado grande'.);
PARAR.

Paso 3 Tome $SUMA = 0$;
$BANDERA1 = 1$; (*Se están sumando los números pequeños.*)
$BANDERA2 = 0$;
$BANDERA3 = 0$;
$HECHO = 0$;
$i = 1$.

Paso 4 Mientras ($i \leq n$ y $BANDERA1 = 1$) haga Paso 5.

Paso 5 Si $|x_i| < y$ entonces establezca $SUMA = SUMA + (Sx_i)^2$;
$i = i + 1$
si no establezca $BANDERA1 = 0$. (*Se halló un número que no es tan pequeño.*)

Paso 6 Si $i > n$ entonces establezca $NORMA = (SUMA)^{1/2}/S$;
$HECHO = 1$
si no establezca $SUMA = (SUMA/S)/S$; (*Cambia la escala de números grandes.*)
$BANDERA2 = 1$.

Paso 7 Mientras ($i \leq n$ y $BANDERA2 = 1$) efectúe Paso 8. (*Suma de números medianos.*)

Paso 8 Si $|x_i| < Y$ entonces establezca $SUMA = SUMA + x_i^2$;
$i = i + 1$
si no establezca $BANDERA2 = 0$. (*Se ha encontrado un número grande.*)

Paso 9 Si $HECHO = 0$ entonces
si $i > n$ entonces establezca $NORMA = (SUMA)^{1/2}$;
$HECHO = 1$
o bien establezca $SUMA = ((SUMA)/s)/s$; (*Cambia de escala a los números grandes.*)
$BANDERA3 = 1$.

Paso 10 Mientras ($i \leq n$ y $BANDERA3 = 1$) ejecute Paso 11.

Paso 11 Establezca $SUMA = SUMA + (sx_i)^2$; (*Suma los números grandes.*)
$i = i + 1$.

Paso 12 Si $HECHO = 0$ entonces
si $SUMA^{1/2} < \lambda s$ entonces establezca $NORMA = (SUMA)^{1/2}/s$;
$HECHO = 1$
o bien establezca $SUMA = \lambda$. (*La norma es demasiado grande.*)

Paso 13 Si $HECHO = 1$ entonces SALIDA ('La norma es', $NORMA$)
o bien SALIDA ('Norma ≥', $NORMA$, 'ocurrió un desbordamiento').

Paso 14 PARAR.

Las relaciones entre las características de la máquina t, σ, λ, *emín*, *emáx* y los parámetros del algoritmo N, s, S, y y Y fueron elegidas en [Brow W, p. 471] como:

$N = 10^{e_N}$, donde $e_N = \lfloor (t-2)/2 \rfloor$, el mayor entero menor o igual a $(t-2)/2$;

$s = 10^{e_s}$, donde $e_s = \lfloor -(emáx + e_N)/2 \rfloor$;

$S = 10^{e_S}$, donde $e_S = \lceil (1 - emín)/2 \rceil$, el menor entero mayor o igual a $(1 - emín)/2$;

$y = 10^{e_y}$, donde $e_y = \lceil (emín + t - 2)/2 \rceil$;

$Y = 10^{e_Y}$, donde $e_Y = \lfloor (emáx - e_N)/2 \rfloor$.

La confiabilidad incorporada a este algoritmo de propósito general ha incrementado en gran medida el grado de dificultad en comparación con el algoritmo de propósito específico mostrado antes en esta sección.

Existen muchos tipos de software comerciales disponibles y de dominio público de propósito general para el análisis numérico. La mayor parte de los primeros software fueron escritos para supercomputadoras (*mainframe*); una buena referencia para esto es *Sources and Developments of Mathematical Software*, editado por Wayne Cowell [Co]. Ahora que la computadora de escritorio se ha vuelto bastante poderosa, se dispone de software numérico común para las computadoras personales y estaciones de trabajo. La mayor parte de este software está escrito en FORTRAN, aunque algunos paquetes están escritos en C, C++ y FORTRAN90.

En 1971 [WR] presentó algunos procedimientos en ALGOL para el cálculo de matrices. Luego, a partir de un paquete de subrutinas de FORTRAN basadas en los procedimientos de ALGOL, se obtuvieron las rutinas EISPACK. Éstas están documentadas en los manuales publicados por Springer-Verlag como parte de su serie *Lecture Notes in Computer*

Sciences [Sm, B] y [Gar]. Las subrutinas en FORTRAN se usan para calcular valores y vectores característicos para una amplia variedad de matrices. El proyecto EISPACK fue el primer paquete de software numérico a gran escala que estuvo disponible para el dominio público y fue la guía de muchos paquetes posteriores. EISPACK recibe mantenimiento de netlib y aparece en la dirección http://www.netlib.org/eispack.

LINPACK es un paquete de subrutinas de FORTRAN para analizar y resolver sistemas de ecuaciones lineales y problemas lineales de mínimos cuadrados. La documentación de este paquete está en [DBMS] y aparece en http://www.netlib.org/linpack. [CV] da una introducción paso a paso a LINPACK, EISPACK y BLAS (por sus siglas en inglés de, subrutinas básicas de álgebra lineal, Basic Linear Algebra Subprograms).

El paquete LAPACK, lanzado al mercado en 1992, es una biblioteca de subrutinas de FORTRAN superior a LINPACK y EISPACK, que integra estos dos conjuntos de algoritmos en un paquete unificado y actualizado. El software se reestructuró para lograr mayor eficiencia con procesadores vectoriales y otros multiprocesadores de alto rendimiento o de memoria compartida. LAPACK se amplió en profundidad y alcance en la versión 3.0, disponible en FORTRAN, FORTRAN90, C, C++ y JAVA. FORTRAN90, C y JAVA sólo están disponibles como interfaces de lenguaje o traducciones de las bibliotecas en FORTRAN de LAPACK. El paquete BLAS no forma parte de LAPACK, pero el código de BLAS se distribuye con LAPACK. El texto *LAPACK User's Guide*, tercera edición [An] está disponible en SIAM o en http://www.netlib.org/lapack/lug/lapack_lug.html. Todo LAPACK o algunas de sus rutinas individuales pueden obtenerse por medio de netlib en netlibornl.gov, netlibresearch.att.com, o http://www.netlib.org/lapack.

Existen otros paquetes de dominio público que sirven para resolver ciertos tipos de problemas. La información acerca de estos programas se puede obtener por medio de correo electrónico, enviando el mensaje "help" a alguna de las siguientes direcciones: netlibresearch.att.com, netlibornl.gov, netlibnac.no o netlibdraci.cs.uow.edu.au o a la dirección uucp adress uunet!research!netlib. Como alternativa a netlib, puede usar Xnetlib para buscar en la base de datos y obtener el software. Hay más información en el artículo *Software Distribution using Netlib*, de Dongarra, Roman y Wade [DRW].

Estos paquetes son muy eficaces, exactos y confiables. Además, se han probado en forma exhaustiva y su documentación es fácil de consultar. Aunque estos programas son portátiles, es bueno investigar su compatibilidad con la máquina y leer con detenimiento toda la documentación. Los programas verifican casi todas las contingencias particulares que podrían producir errores y fallas. Al final de cada capítulo analizaremos algunos de los paquetes adecuados de propósito general.

Los paquetes comerciales también representan lo más avanzado en métodos numéricos. Por lo general, su contenido se basa en los paquetes de dominio público, pero incluyen técnicas para casi cualquier tipo de problema.

IMSL (del inglés International Mathematical Software Library, bibliotecas internacionales en matemáticas y estadística) incluye las bibliotecas MATH, STAT y SFUN para matemáticas numéricas, estadística y funciones especiales del análisis numérico, respectivamente. Estas bibliotecas contienen más de 900 subrutinas disponibles originalmente en FORTRAN 77 y ahora en C++, FORTRAN90 y JAVA. Con estas subrutinas se resuelven los problemas más comunes de análisis numérico. En 1970, IMSL se convirtió en la primera biblioteca científica de gran escala para las supercomputadoras (*mainframe*). Desde entonces, las bibliotecas han estado disponibles para sistemas de cómputo que van desde las supercomputadoras hasta las computadoras personales. Éstas pueden adquirirse en Visual Numerics, 9990 Richmond Ave S400, Houston, TX 77042−4548, con dirección en Internet http://www.vni.com. Los paquetes se entregan en forma compilada, con amplia

documentación. Hay un programa ejemplo para cada rutina, así como información básica de referencia. Las IMSL contienen métodos para sistemas lineales, análisis de sistemas característicos, interpolación y aproximación, integración y derivación, ecuaciones diferenciales, transformadas, ecuaciones no lineales, optimización y operaciones básicas con matrices y vectores. La biblioteca contiene además muchas rutinas de estadística.

El Numerical Algorithms Group (NAG) se fundó en 1970, en Reino Unido. NAG ofrece más de 1000 subrutinas en una biblioteca de FORTRAN 77, cerca de 400 subrutinas en una biblioteca en C, más de 200 subrutinas en su biblioteca de FORTRAN 90 y una biblioteca numérica de FORTRAN MPI para máquinas en paralelo y cúmulos de estaciones de trabajo o computadoras personales. Un subconjunto de su biblioteca en FORTRAN 77 (la NAG Foundation Library) está disponible para computadoras personales y estaciones de trabajo donde el espacio está limitado. Las bibliotecas NAG C, FORTRAN 90 y FORTRAN MPI ofrecen muchas de las mismas rutinas de la biblioteca FORTRAN. El manual del usuario de NAG incluye instrucciones y ejemplos, junto con una salida muestra para cada una de las rutinas. [Ph] es una introducción útil para las rutinas NAG. La biblioteca NAG contiene rutinas que permiten realizar la mayor parte de las tareas estándar de análisis numérico de manera similar a las IMSL. También incluye algunas rutinas de estadística y un conjunto de rutinas gráficas. La biblioteca se puede comprar en Numerical Algorithms Group, Inc., 1400 Opus Place, Suite 200, Downers Grove, IL 60515-5702, con dirección en Internet http://www.nag.com.

Los paquetes IMSL y NAG están diseñados para los matemáticos, científicos e ingenieros que desean llamar desde un programa las subrutinas FORTRAN de alta calidad dentro de un programa. La documentación que incluyen los paquetes comerciales explica el programa maestro que se necesita para usar las rutinas de la biblioteca. Los siguientes tres paquetes de software son ambientes independientes. Al activarlos, el usuario teclea instrucciones para que el paquete resuelva el problema. Sin embargo, cada paquete permite programar dentro del lenguaje de sus instrucciones.

MATLAB es un laboratorio de matrices que originalmente era un programa en FORTRAN publicado por Cleve Moler [Mo]. El laboratorio se basa sobre todo en las subrutinas EISPACK y LINPACK, aunque se han incorporado funciones como sistemas lineales, integración numérica, trazadores cúbicos, ajuste de curvas, optimización, ecuaciones diferenciales ordinarias y herramientas gráficas. MATLAB está escrito actualmente en C y en lenguaje ensamblador, y la versión para computadora personal requiere un coprocesador numérico. La estructura básica consiste en realizar operaciones con matrices, como determinar los valores característicos de una matriz introducida desde la línea de instrucciones o desde un archivo externo mediante llamadas de funciones. Éste es un poderoso sistema autosuficiente que resulta muy útil para la enseñanza de álgebra lineal aplicada. MATLAB está en el mercado desde 1985 y puede adquirirse en The Math Works, Inc., Cochituate Place, 24 Prime Park Way, Natick, MA 01760. La dirección de correo electrónico de The Math Works es infomathworks.com y la dirección en Internet es http://www.mathworks.com. El software de MATLAB está diseñado para ejecutarse en muchas computadoras, incluyendo las computadoras personales compatibles con IBM, APPLE Macintosh y las estaciones de trabajo SUN. La versión de MATLAB para el estudiante no requiere un coprocesador pero lo usará en caso que esté disponible.

El segundo paquete es GAUSS, un sistema matemático y estadístico producido por Lee E. Ediefson y Samuel D. Jones en 1985. Está codificado en lenguaje ensamblador y se basa en EISPACK y LINPACK. Como en el caso de MATLAB, dispone de integración, derivación, sistemas no lineales, transformadas rápidas de Fourier y gráficas. GAUSS está menos orientado a la enseñanza en álgebra lineal y más hacia el análisis estadístico de

datos. Este paquete también usa un coprocesador numérico si está disponible. Puede adquirirse en Aptech Systems, Inc., 23804 S.E. Kent-Kangley Road, Maple Valley, WA 98038 (infoaptech.com).

El tercer paquete es Maple, un sistema de álgebra por computadora desarrollado en 1980 por The Symbolic Computational Group de la Universidad de Waterloo. El diseño del sistema Maple original se presenta en el artículo de B.W. Char, K.O. Geddes, W.M. Gentlemen y G.H. Gonnet [CGGG]. Maple está en el mercado desde 1985 y se consigue en Waterloo Maple Inc., 57 Erb Street, Waterloo ON N2L 6C2. La dirección de correo electrónico de Waterloo Maple es infomaplesoft.com y la dirección en Internet es http://www.maplesoft.com Maple está escrito en C y tiene la capacidad de manejar la información de manera simbólica, lo cual permite al usuario obtener respuestas exactas en vez de valores numéricos. Con Maple se pueden obtener respuestas exactas a problemas matemáticos como integrales, ecuaciones diferenciales y sistemas lineales. Contiene una estructura de programación y permite guardar texto e instrucciones en sus archivos de hoja de trabajo. Luego, estas hojas de trabajo se pueden introducir en Maple y ejecutar las instrucciones. Debido a sus características de cálculo simbólico, cálculo numérico y hojas de trabajo, se eligió a Maple como lenguaje para este texto. En todo el libro se intercalan instrucciones de Maple.

Existen muchos paquetes que pueden ser catalogados como paquetes de supercalculadora para computadoras personales. Sin embargo, estos no deben confundirse con el software de propósito general mencionado aquí. Si usted tiene interés en alguno de estos paquetes, debe leer *Supercalculators on the PC* por B. Simon y R.M. Wilson [SW].

Para obtener más información acerca del software y las bibliotecas de software consulte los libros de Cody y Waite [CW] y Kockler [Ko]; así como en el artículo de 1995 de Dongarra y Walker [DW]. En el libro de Chaitini-Chatelin y Frayse [CF] y el artículo de Goldberg [Go] hay más información acerca del cálculo de punto flotante.

Algunos libros dedicados a la aplicación de técnicas numéricas en computadoras en paralelo son Schendell [Sche], Phillips y Freeman [PF] y Golub y Ortega [GO].

CAPÍTULO

2

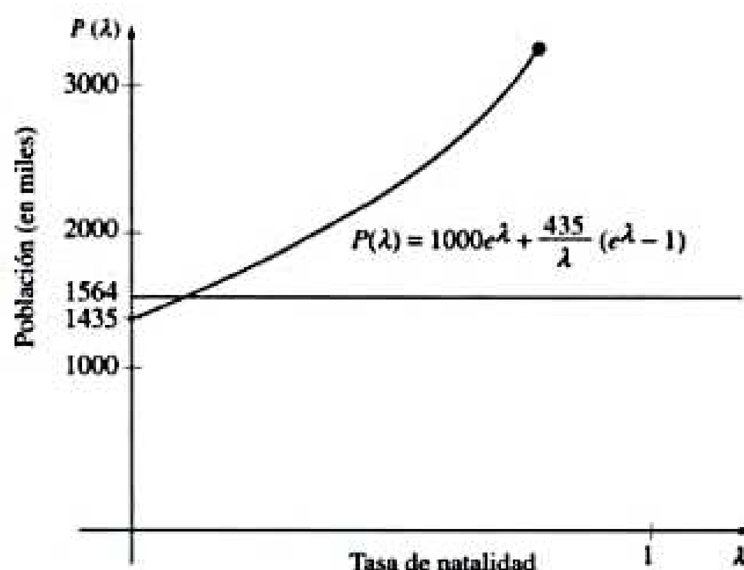
Soluciones de ecuaciones de una variable

. . .

El crecimiento de una población numerosa puede modelarse durante periodos breves, con sólo suponer que ésta crece constantemente con el tiempo a una tasa que es proporcional al número de habitantes que existen en ese tiempo. Si denotamos con $N(t)$ la cantidad de habitantes en el tiempo t y con λ el índice constante de natalidad, la población satisface la ecuación diferencial

$$\frac{dN(t)}{dt} = \lambda N(t).$$

La solución de esta ecuación es $N(t) = N_0 e^{\lambda t}$, donde N_0 denota la población inicial.



Este modelo exponencial es válido sólo cuando la población se halla aislada, es decir, sin que exista inmigración proveniente del exterior. Si se permite la inmigración con una tasa constante v la ecuación diferencial que rige la situación será

$$\frac{dN(t)}{dt} = \lambda N(t) + v,$$

cuya solución es

$$N(t) = N_0 e^{\lambda t} + \frac{v}{\lambda}(e^{\lambda t} - 1).$$

Supóngase que cierta población tiene inicialmente un millón de habitantes, que 435 000 de ellos inmigran hacia la comunidad durante el primer año y que 1 564 000 se encuentran en ella al final del año 1. Si queremos determinar la natalidad de esta población, debemos determinar λ en la ecuación

$$1\,564\,000 = 1\,000\,000 e^{\lambda} + \frac{435\,000}{\lambda}(e^{\lambda} - 1).$$

Los métodos numéricos que se tratan en este capítulo sirven para obtener aproximaciones a las soluciones de este tipo de ecuaciones, cuando no es posible obtener respuestas exactas con métodos algebraicos. En el ejercicio 20 de la sección 2.3 se considera la solución de este problema en particular.

2.1 El método de bisección

En este capítulo estudiaremos uno de los problemas básicos de la aproximación numérica: el *problema de la búsqueda de raíces*. Consiste en obtener una *raíz*, o solución, de una ecuación de la forma $f(x) = 0$ para una función dada f. (Al número x se le llama también *cero* de f.) El problema de encontrar una aproximación a la raíz de una ecuación se remonta por lo menos al año 1700 a.C. Una tabla cuneiforme que pertenece a la Yale Babylonian Collection, y que data de este periodo, da un número sexagesimal (base 60) equivalente a 1.414222 como aproximación a $\sqrt{2}$, resultado que tiene una precisión de hasta 10^{-5}. Esta aproximación se puede determinar mediante una técnica descrita en el ejercicio 19 de la sección 2.2.

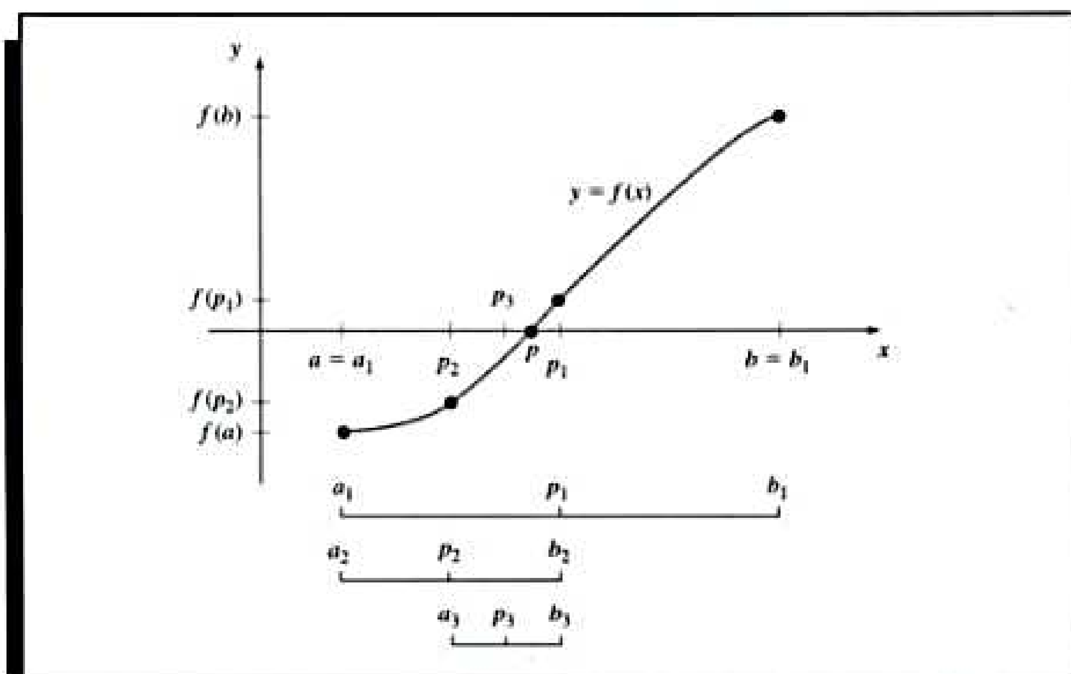
La primera técnica, que se basa en el teorema del valor intermedio, se conoce con el nombre de **método de bisección** o de **búsqueda binaria**. Supongamos que f es una función continua definida en el intervalo $[a, b]$ con $f(a)$ y $f(b)$ de signos diferentes. De acuerdo con el teorema del valor intermedio, existe un número p en (a, b) tal que $f(p) = 0$. Si bien el procedimiento se aplica aunque exista más de una raíz en el intervalo (a, b), por razones de simplicidad suponemos que la raíz de este intervalo es única. El método requiere dividir varias veces a la mitad los subintervalos de $[a, b]$ y, en cada paso, localizar la mitad que contenga a p.

Para empezar, supongamos que $a_1 = a$ y $b_1 = b$, y sea p_1 el punto medio de $[a, b]$; es decir,

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}.$$

Si $f(p_1) = 0$, entonces $p = p_1$; de no ser así, entonces $f(p_1)$ tiene el mismo signo que $f(a_1)$ o $f(b_1)$. Si $f(p_1)$ y $f(a_1)$ tienen el mismo signo, entonces $p \in (p_1, b_1)$ y tomamos $a_2 = p_1$ y $b_2 = b_1$. Si $f(p_1)$ y $f(a_1)$ tienen signos opuestos, entonces $p \in (a_1, p_1)$ y tomamos $a_2 = a_1$ y $b_2 = p_1$. Después volvemos a aplicar el proceso al intervalo $[a_2, b_2]$. Esto nos da el método que se describe en el algoritmo 2.1 (Véase fig. 2.1).

Figura 2.1



ALGORITMO 2.1

Bisección

Para obtener una solución a $f(x) = 0$ dada la función f continua en el intervalo $[a, b]$, donde $f(a)$ y $f(b)$ tienen signos opuestos:

ENTRADA extremos a, b; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de error.

Paso 1 Tome $i = 1$;
$FA = f(a)$.

Paso 2 Mientras $i \leq N_0$ haga pasos 3-6.

Paso 3 Tome $p = a + (b - a)/2$; (*Calcule* p_i.)
$FP = f(p)$.

Paso 4 Si $FP = 0$ o $(b - a)/2 < TOL$ entonces
SALIDA (p); (*Procedimiento terminado satisfactoriamente.*)
PARAR.

Paso 5 Tome $i = i + 1$.

Paso 6 Si $FA \cdot FP > 0$ entonces tome $a = p$; (Calcule a_i, b_i.)
$FA = FP$
si no tome $b = p$.

Paso 7 SALIDA ('El método fracasó después de N_0, iteraciones, N_0 = ', N_0);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

A continuación describiremos otros procedimientos de paro que pueden aplicarse en el paso 4 del algoritmo 2.1 o a cualquiera de las técnicas iterativas que se estudian en este capítulo. Por ejemplo, seleccione una tolerancia $\epsilon > 0$ y genere $p_1, \ldots, p_N$ hasta que se satisfaga una de las siguientes condiciones:

$$|p_N - p_{N-1}| < \epsilon, \tag{2.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \epsilon, \quad p_N \neq 0, \quad \text{o} \tag{2.2}$$

$$|f(p_N)| < \epsilon. \tag{2.3}$$

Por desgracia, al usar cualquiera de estos criterios de paro pueden surgir problemas. Por ejemplo, existen sucesiones $\{p_n\}_{n=0}^{\infty}$ con la propiedad de que las diferencias $p_n - p_{n-1}$ convergen a cero, mientras que la sucesión diverge (ejercicio 15). También es posible que $f(p_n)$ esté cercano a cero, mientras que p_n difiere significativamente de p (véase el ejercicio 14.) En caso de que no se conozcan f o p, la desigualdad (2.2) es el mejor criterio de paro al que puede recurrirse, ya que verifica el error relativo.

Cuando se generan aproximaciones por medio de la computadora, conviene fijar el número máximo de iteraciones que se efectuarán; así se evitará introducir un ciclo infinito, posibilidad que puede presentarse cuando la sucesión diverge (y también cuando el programa no está codificado correctamente). Esto se hace en el paso 2 del algoritmo 2.1, donde se estableció la frontera inicial N_0 y el procedimiento terminaba si $i > N_0$.

Obsérvese que para iniciar el algoritmo de bisección, hay que encontrar un intervalo $[a, b]$, de modo que $f(a) \cdot f(b) < 0$. En cada paso, la longitud del intervalo que se sabe que contiene un cero de f se reduce en un factor de 2; por tanto, conviene escoger un intervalo inicial $[a, b]$ lo más pequeño posible. Por ejemplo, si $f(x) = 2x^3 - x^2 + x - 1$, entonces

$$f(-4) \cdot f(4) < 0 \qquad \text{y} \qquad f(0) \cdot f(1) < 0,$$

de manera que el algoritmo de bisección pueda emplearse en uno de los intervalos $[-4, 4]$ o $[0, 1]$. Al comenzar el algoritmo de bisección en $[0, 1]$ y no en $[-4, 4]$, la cantidad de iteraciones necesarias para alcanzar determinada exactitud disminuirá en 3.

El siguiente ejemplo ilustra el algoritmo de bisección. En este ejemplo la iteración se termina cuando el error relativo es menor que 0.0001, es decir, cuando

$$\frac{|p - p_n|}{|p|} < 10^{-4}.$$

EJEMPLO 1 La ecuación $f(x) = x^3 + 4x^2 - 10 = 0$ tiene una raíz en [1, 2], ya que $f(1) = -5$ y $f(2) = 14$. El algoritmo de bisección da los valores de la tabla 2.1.

Tabla 2.1

n	a_n	b_n	p_n	$f(p_n)$
1	1.0	2.0	1.5	2.375
2	1.0	1.5	1.25	−1.79687
3	1.25	1.5	1.375	0.16211
4	1.25	1.375	1.3125	−0.84839
5	1.3125	1.375	1.34375	−0.35098
6	1.34375	1.375	1.359375	−0.09641
7	1.359375	1.375	1.3671875	0.03236
8	1.359375	1.3671875	1.36328125	−0.03215
9	1.36328125	1.3671875	1.365234375	0.000072
10	1.36328125	1.365234375	1.364257813	−0.01605
11	1.364257813	1.365234375	1.364746094	−0.00799
12	1.364746094	1.365234375	1.364990235	−0.00396
13	1.364990235	1.365234375	1.365112305	−0.00194

Después de 13 iteraciones, $p_{13} = 1.365112305$ aproxima la raíz p con un error de

$$|p - p_{13}| < |b_{14} - a_{14}| = |1.365234375 - 1.365112305| = 0.000122070.$$

Puesto que $|a_{14}| < |p|$,

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9.0 \times 10^{-5},$$

la aproximación será correcta al menos en cuatro dígitos significativos. El valor correcto de p, con nueve cifras decimales, es $p = 1.365230013$. Observe que p_9 está más cerca de p que la aproximación final p_{13}. Podemos suponer que esto es verdad ya que $|f(p_9)| < |f(p_{13})|$, pero no podemos verificarlo si no conocemos la respuesta verdadera. ■

El método de bisección, aunque claro desde el punto de vista conceptual, ofrece inconvenientes importantes, como el de converger lentamente (es decir, N puede volverse muy grande antes que $|p - p_N|$ sea bastante pequeño) e inadvertidamente podemos desechar una buena aproximación intermedia. Sin embargo, tiene la importante propiedad de que siempre converge en una solución y, por tal razón, a menudo sirve para iniciar los métodos más eficientes que explicaremos más adelante en el capítulo.

Teorema 2.1 Supongamos que $f \in C[a, b]$ y $f(a) \cdot f(b) < 0$. El método de bisección que se usa en el algoritmo 2.1 genera una sucesión $\{p_n\}_{n=1}^{\infty}$ que aproxima a un cero de p de f, tal que

$$|p_n - p| \le \frac{b-a}{2^n}, \quad \text{donde} \quad n \ge 1. \qquad \blacksquare$$

Demostración Para cada $n \ge 1$, tenemos

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \quad \text{y} \quad p \in (a_n, b_n).$$

Y como $p_n = \frac{1}{2}(a_n + b_n)$ para toda $n \ge 1$, se deduce que

$$|p_n - p| \le \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}.$$

■ ■ ■

Ya que

$$|p_n - p| \le (b-a)\frac{1}{2^n},$$

esta desigualdad implica que $\{p_n\}_{n=1}^{\infty}$ converge a p con una razón de convergencia $O(\frac{1}{2^n})$; es decir,

$$p_n = p + O\left(\frac{1}{2^n}\right).$$

Es importante señalar que el teorema 2.1 da sólo una cota del error de aproximación y que ésta puede ser extremadamente conservadora. Por ejemplo, cuando la aplicamos al problema del ejemplo 1 sólo garantiza que

$$|p - p_9| \le \frac{2-1}{2^9} \approx 2 \times 10^{-3},$$

pero el error real es mucho menor:

$$|p - p_9| = |1.365230013 - 1.365234375| \approx 4.4 \times 10^{-6}.$$

EJEMPLO 2 Para determinar la cantidad de iteraciones necesarias para resolver $f(x) = x^3 + 4x^2 - 10 = 0$ con una exactitud de 10^{-3} por medio de $a_1 = 1$ y de $b_1 = 2$ hay que encontrar un entero N que satisfaga

$$|p_N - p| \le 2^{-N}(b-a) = 2^{-N} < 10^{-3}.$$

Para determinar N usaremos logaritmos. Aunque se podrían usar logaritmos de cualquier base, utilizaremos los de base 10, porque la tolerancia está dada como una potencia de 10. Puesto que $2^{-N} < 10^{-3}$ implica que $\log_{10} 2^{-N} < \log_{10} 10^{-3} = -3$, tendremos

$$-N \log_{10} 2 < -3 \quad \text{y} \quad N > \frac{3}{\log_{10} 2} \approx 9.96.$$

Por tanto, se necesitan unas diez iteraciones para lograr una aproximación exacta dentro de 10^{-3}. La tabla 2.1 muestra que el valor de $p_9 = 1.365234375$ es exacto dentro de 10^{-4}.

Conviene recordar que el análisis de error no da más que una cota del número de iteraciones necesarias; muchas veces esta cota es mucho mayor que el número que se requiere. ■

La cota para el número de iteraciones en el método de bisección supone que los cálculos se realizan en una aritmética con una infinidad de dígitos. Al aplicarse el método en una computadora, hay que tomar en cuenta los efectos de los errores por redondeo. Por ejemplo, el cálculo del punto medio del intervalo $[a_n, b_n]$ debe encontrarse mediante la ecuación

$$p_n = a_n + \frac{b_n - a_n}{2}$$

y no con la ecuación algebraicamente equivalente

$$p_n = \frac{a_n + b_n}{2}.$$

La primera ecuación agrega una pequeña correción, $(b_n - a_n)/2$, al valor conocido a_n. Cuando $b_n - a_n$ está cerca de la precisión máxima de la máquina, esta corrección podría tener un error, pero éste no afectaría de manera significativa el valor calculado de p_n. Sin embargo, cuando $b_n - a_n$ está cerca de la precisión máxima de la máquina, es posible que $(a_n + b_n)/2$ regrese un punto medio que ni siquiera esté en el intervalo $[a_n, b_n]$.

Como observación final, para determinar cuál subintervalo de $[a_n, b_n]$ contiene una raíz de f, es mejor usar la función **signo**, que se define como

$$\text{signo}\,(x) = \begin{cases} -1, & \text{si } x < 0, \\ 0, & \text{si } x = 0, \\ 1, & \text{si } x > 0. \end{cases}$$

El criterio

$$\text{signo}\,(f(a_n))\ \text{signo}\,(f(p_n)) < 0 \qquad \text{en lugar de} \qquad f(a_n)\,f(p_n) < 0$$

da el mismo resultado pero evita la posibilidad de un sobreflujo en la multiplicación de $f(a_n)$ y $f(p_n)$.

CONJUNTO DE EJERCICIOS 2.1

1. Aplique el método de bisección para obtener p_3 para $f(x) = \sqrt{x} - \cos x$ en $[0, 1]$.

2. Sea $f(x) = 3(x + 1)\,(x - \frac{1}{2})\,(x - 1)$. Aplique el método de bisección a los siguientes intervalos para encontrar p_3.

a. $[-2, 1.5]$ **b.** $[-1.25, 2.5]$

3. Aplique el método de bisección para encontrar las soluciones exactas dentro de 10^{-2} para $x^3 - 7x^2 + 14x - 6 = 0$ en cada intervalo.

a. $[0, 1]$ **b.** $[1, 3.2]$ **c.** $[3.2, 4]$

4. Aplique el método de bisección para encontrar las soluciones exactas dentro de 10^{-2} para $x^4 - 2x^3 - 4x^2 + 4x + 4 = 0$ en cada intervalo.

a. $[-2, -1]$ **b.** $[0, 2]$ **c.** $[2, 3]$ **d.** $[-1, 0]$

5. Use el método de bisección para encontrar una solución exacta dentro de 10^{-3} para $x = \tan x$ en $[4, 4.5]$.

6. Use el método de bisección para encontrar una solución exacta dentro de 10^{-3} para $2 + \cos(e^x - 2) - e^x = 0$ en $[0.5, 1.5]$.

7. Aplique el método de bisección para encontrar soluciones exactas dentro de 10^{-5} para los siguientes problemas.

a. $x - 2^{-x} = 0$ para $0 \leq x \leq 1$

b. $e^x - x^2 + 3x - 2 = 0$ para $0 \leq x \leq 1$

c. $2x\cos(2x) - (x + 1)^2 = 0$ para $-3 \leq x \leq -2$ y para $-1 \leq x \leq 0$

d. $x\cos x - 2x^2 + 3x - 1 = 0$ para $0.2 \leq x \leq 0.3$ y para $1.2 \leq x \leq 1.3$

8. Sea $f(x) = (x + 2)(x + 1)^2 x(x - 1)^3(x - 2)$. ¿A cuál cero de f converge el método de bisección en los siguientes intervalos?

a. $[-1.5, 2.5]$ **b.** $[-0.5, 2.4]$ **c.** $[-0.5, 3]$ **d.** $[-3, -0.5]$

9. Sea $f(x) = (x + 2)(x + 1)x(x - 1)^3(x - 2)$. ¿A cuál cero de f converge el método de bisección en los siguientes intervalos?

a. $[-3, 2.5]$ **b.** $[-2.5, 3]$ **c.** $[-1.75, 1.5]$ **d.** $[-1.5, 1.75]$

10. Encuentre una aproximación a $\sqrt{3}$ correcta con una exactitud de 10^{-4} usando el algoritmo de bisección. [*Sugerencia*: considere $f(x) = x^2 - 3$.]

11. Encuentre una aproximación a $\sqrt[3]{25}$ correcta en 10^{-4} por medio del algoritmo de bisección.

12. Use el teorema 2.1 para obtener una cota del número de iteraciones que se requieren para alcanzar una aproximación con una exactitud de 10^{-3} a la solución de $x^3 + x - 4 = 0$ que se encuentra en el intervalo $[1, 4]$. Obtenga una aproximación de la raíz con este grado de exactitud.

13. Use el teorema 2.1 para obtener una cota del número de iteraciones que se requieren para alcanzar una aproximación con una exactitud de 10^{-4} a la solución de $x^3 - x - 1 = 0$ que se encuentra en el intervalo $[1, 2]$. Obtenga una aproximación de la raíz con este grado de exactitud.

14. Sea $f(x) = (x - 1)^{10}$, $p = 1$, y $p_n = 1 + 1/n$. Demuestre que $|f(p_n)| < 10^{-3}$ siempre que $n > 1$, pero que $|p - p_n| < 10^{-3}$ requiere que $n > 1000$.

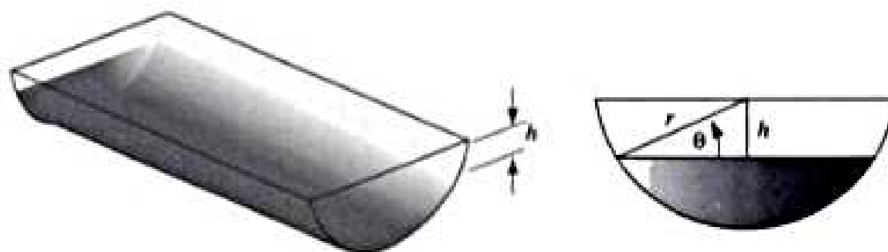
15. Sea $\{p_n\}$ la sucesión definida por $p_n = \sum_{k=1}^{n} \frac{1}{k}$. Demuestre que $\{p_n\}$ diverge aun cuando $\lim_{n\to\infty}(p_n - p_{n-1}) = 0$.

16. La función definida por $f(x) = \operatorname{sen} \pi x$ tiene ceros en todos los enteros. Muestre que cuando $-1 < a < 0$ y $2 < b < 3$, el método de bisección converge a

a. 0, si $a + b < 2$ **b.** 2, si $a + b > 2$ **c.** 1, si $a + b = 2$

17. Un abrevadero de longitud L tiene una sección transversal en forma de semicírculo con radio r (Véase la figura anexa.) Cuando se llena de agua hasta una distancia h de la parte superior, el volumen V de agua es

$$V = L\,[0.5\pi r^2 - r^2 \operatorname{arcsen}(h/r) - h(r^2 - h^2)^{1/2}].$$



Suponga que $L = 10$ pies, $r = 1$ pie, y que $V = 12.4$ pies^3. Determine la profundidad del agua en el abrevadero hasta 0.01 pies.

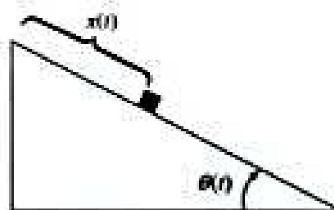
18. Una partícula parte del reposo sobre un plano inclinado uniforme, cuyo ángulo θ cambia con una rapidez constante de

$$\frac{d\theta}{dt} = \omega < 0.$$

Al final de t segundos, la posición del objeto está dada por

$$x(t) = -\frac{g}{2\omega^2}\left(\frac{e^{\omega t} - e^{-\omega t}}{2} - \operatorname{sen} \omega t\right).$$

Suponga que la partícula se desplazó 1.7 pies en 1 s. Encuentre, con una exactitud de 10^{-5}, la rapidez ω con que θ cambia. Suponga que $g = 32.17$ pies/s^2.



2.2 Iteración de punto fijo

Un **punto fijo** de una función g es un número p para el cual $g(p) = p$. En esta sección estudiaremos el problema de encontrar las soluciones a los problemas de punto fijo y la conexión entre éstos y los de búsqueda de la raíz que deseamos resolver.

Los problemas de búsqueda de raíces y los de punto fijo son clases equivalentes en el siguiente sentido:

> Dado un problema de buscar una raíz $f(p) = 0$, podemos definir una función g con un punto fijo en p de diversas formas; por ejemplo, como $g(x) = x - f(x)$ o como $g(x) = x + 3f(x)$. Por el contrario, si la función g tiene un punto fijo en p, entonces la función definida por $f(x) = x - g(x)$ tiene un cero en p.

Aunque los problemas que queremos resolver vienen en forma de búsqueda de raíces, la forma de punto fijo es más fácil de analizar; algunas opciones de punto fijo dan origen a técnicas muy poderosas de búsqueda de raíces.

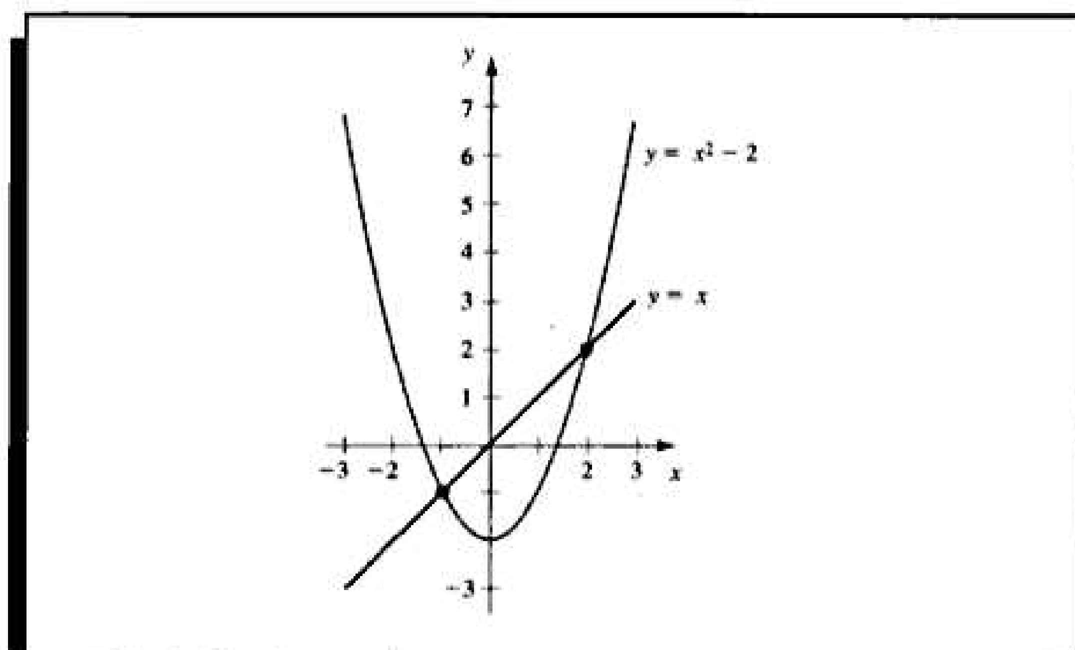
Lo primero que debemos hacer es acostumbrarnos a este nuevo tipo de problema, y decidir cuándo una función tiene un punto fijo y cómo podemos aproximar los puntos fijos con determinado grado de precisión.

EJEMPLO 1 La función $g(x) = x^2 - 2$, para $-2 \le x \le 3$, tiene puntos fijos en $x = -1$ y en $x = 2$, porque

$$g(-1) = (-1)^2 - 2 = -1 \quad \text{y} \quad g(2) = 2^2 - 2 = 2. \qquad \blacksquare$$

Esto podemos observarlo en la figura 2.2.

Figura 2.2



El siguiente teorema contiene suficientes condiciones para la existencia y unicidad del punto fijo.

Teorema 2.2

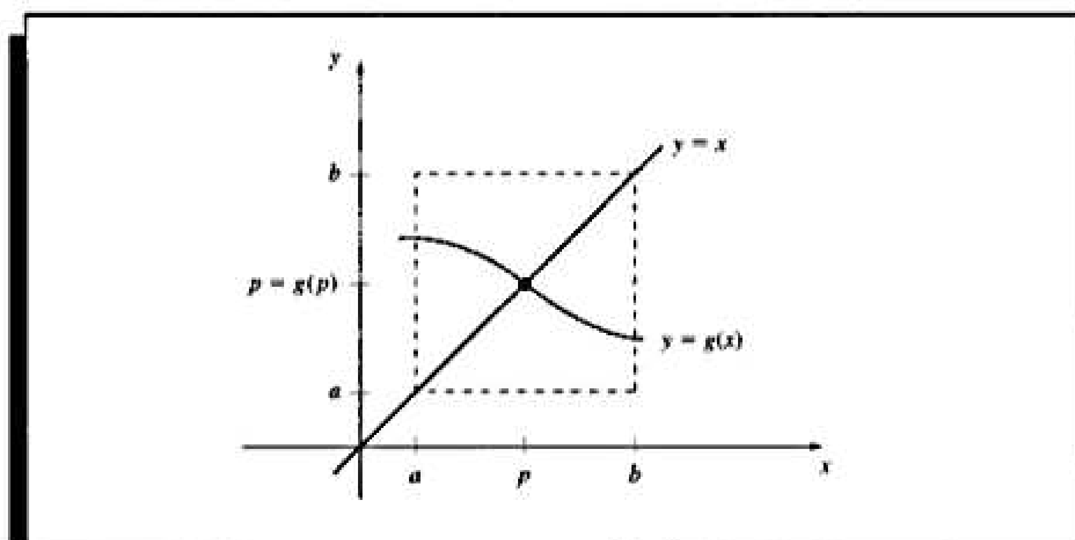
a. Si $g \in C[a, b]$ y $g(x) \in [a, b]$, para toda $x \in [a, b]$, entonces g tiene un punto fijo en $[a, b]$.

b. Y si además $g'(x)$ existe en (a, b) y existe una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \qquad \text{para toda} \qquad x \in (a, b),$$

entonces el punto fijo en $[a, b]$ es único (véase Fig. 2.3). ■

Figura 2.3



Demostración

a. Si $g(a) = a$ o si $g(b) = b$, entonces g tendrá un punto fijo en un extremo. Supongamos que no es así; entonces deberá ser cierto que $g(a) > a$ y que $g(b) < b$. La función $h(x) = g(x) - x$ es continua en $[a, b]$ y tenemos

$$h(a) = g(a) - a > 0 \quad \text{y} \quad h(b) = g(b) - b < 0.$$

El teorema del valor intermedio establece que existe una $p \in (a, b)$ para la cual $h(p) = 0$. Ese número p es un punto fijo de g.

$$0 = h(p) = g(p) - p \quad \text{implica que} \quad g(p) = p.$$

b. Suponga además que $|g'(x)| \leq k < 1$ y que p y q son puntos fijos en $[a, b]$ tal que $p \neq q$. Según el teorema del valor medio, existe un número ξ entre p y q y, por tanto, en $[a, b]$ tal que

$$\frac{g(p) - g(q)}{p - q} = g'(\xi).$$

Por tanto,

$$|p - q| = |g(p) - g(q)| = |g'(\xi)|\,|p - q| \leq k|p - q| < |p - q|,$$

lo cual es una contradicción. Esta contradicción se debe solamente a la suposición, $p \neq q$. Por tanto, $p = q$ y el punto fijo en $[a, b]$ es único. ■ ■ ■

EJEMPLO 2 **a.** Sea $g(x) = (x^2 - 1)/3$ en $[-1, 1]$. El teorema del valor extremo establece que el mínimo absoluto de g ocurre en $x = 0$ y $g(0) = -\frac{1}{3}$. De manera análoga, el máximo absoluto de g ocurre en $x = \pm 1$ y tiene el valor $g(\pm 1) = 0$. Además, g es continua y

$$|g'(x)| = \left|\frac{2x}{3}\right| \leq \frac{2}{3}, \quad \text{para toda } x \in (-1, 1).$$

Por tanto, g satisface todas las hipótesis del teorema 2.2 y tiene un punto fijo único en $[-1, 1]$.

En este ejemplo el único punto fijo p en el intervalo $[-1, 1]$ puede determinarse algebraicamente. Si

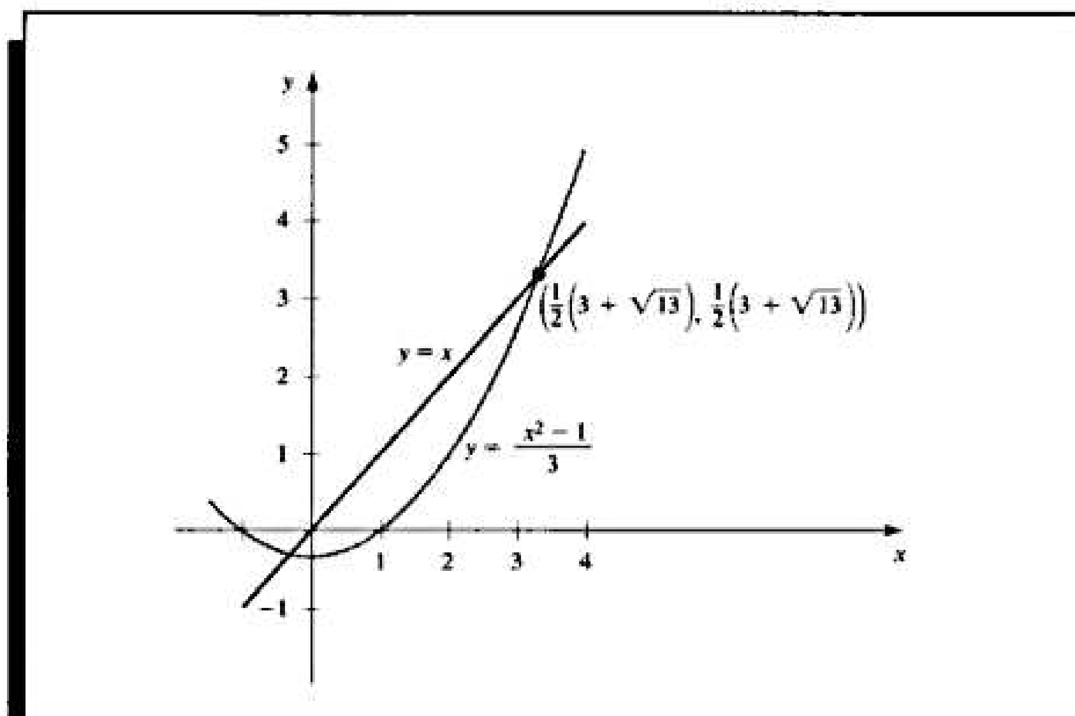
$$p = g(p) = \frac{p^2 - 1}{3}, \quad \text{por tanto} \quad p^2 - 3p - 1 = 0,$$

lo cual, por la fórmula cuadrática, implica que

$$p = \frac{1}{2}(3 - \sqrt{13}).$$

Nótese que g también tiene un punto fijo único $p = \frac{1}{2}(3 + \sqrt{13})$ en el intervalo $[3, 4]$. Sin embargo, $g(4) = 5$ y $g'(4) = \frac{8}{3} > 1$, así que g no satisface las hipótesis del teorema 2.2 en $[3, 4]$. Esto demuestra que esas hipótesis son suficientes para garantizar un punto fijo único, pero no son necesarias (véase Fig. 2.4).

Figura 2.4



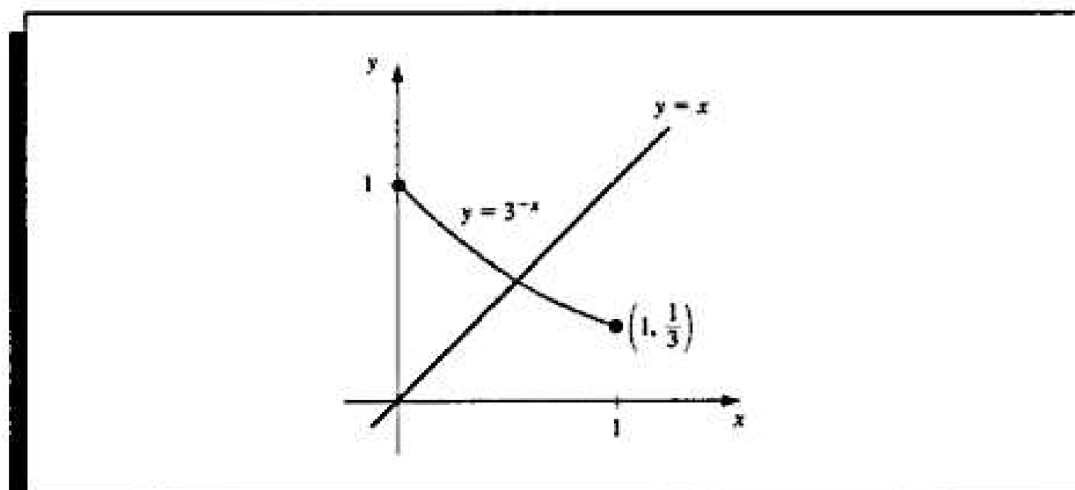
b. Sea $g(x) = 3^{-x}$. Puesto que $g'(x) = -3^{-x} \ln 3 < 0$ en $[0, 1]$, la función g es decreciente en $[0, 1]$. Por tanto

$$g(1) = \tfrac{1}{3} \leq g(x) \leq 1 = g(0), \qquad \text{para} \qquad 0 \leq x \leq 1.$$

Así, para $x \in [0, 1]$, tendremos $g(x) \in [0, 1]$, y g tendrá un punto fijo en $[0, 1]$. Puesto que

$$g'(0) = -\ln 3 = -1.098612289,$$

Figura 2.5



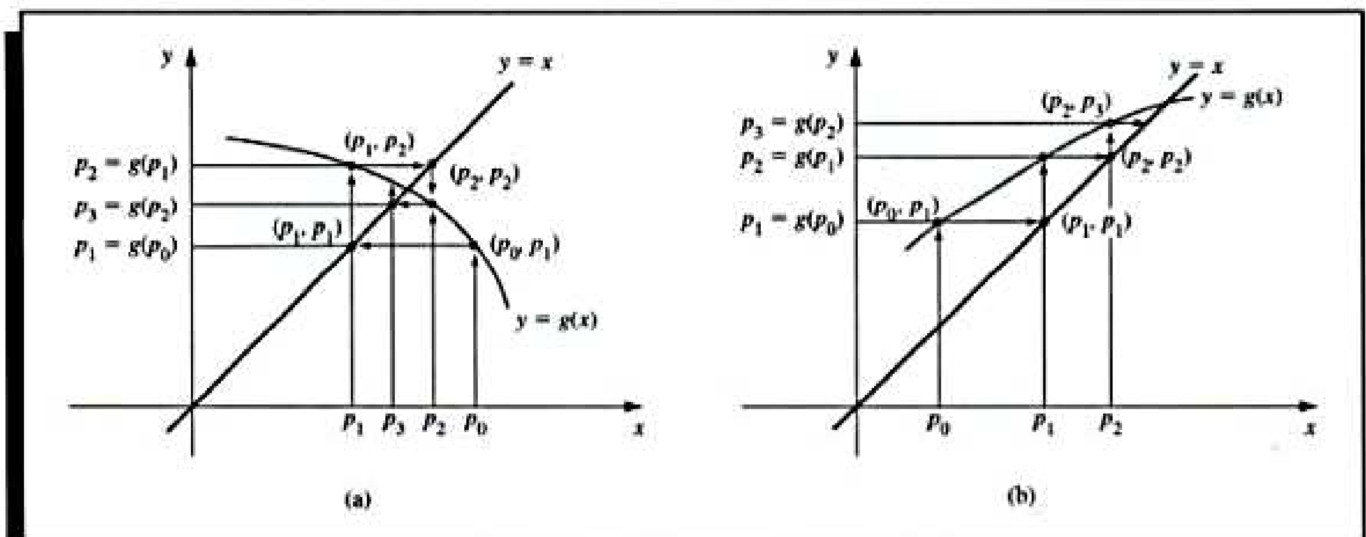
$|g'(x)| \not\leq 1$ en $(0, 1)$, y no se puede utilizar el teorema 2.2 para determinar unicidad. Sin embargo, g siempre es decreciente y en la figura 2.5 se observa claramente que el punto fijo ha de ser único. ■

Para aproximar el punto fijo de una función g, escogemos una aproximación inicial p_0 y generamos la sucesión $\{p_n\}_{n=0}^{\infty}$ haciendo $p_n = g(p_{n-1})$ para cada $n \geq 1$. Si la secuencia converge en p y si g es continua, entonces

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\left(\lim_{n\to\infty} p_{n-1}\right) = g(p),$$

y obtenemos una solución con $x = g(x)$. Esta técnica recibe el nombre de **iteración de punto fijo** o **iteración funcional**. Este procedimiento se describe detalladamente en el algoritmo 2.2 y se muestra gráficamente en la figura 2.6.

Figura 2.6



ALGORITMO 2.2

Iteración de punto fijo

Para obtener una solución a $p = g(p)$ dada una aproximación inicial p_0:

ENTRADA aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de error.

Paso 1 Tome $i = 1$.

Paso 2 Mientras $i \leq N_0$ haga pasos 3-6.

Paso 3 Tome $p = g(p_0)$. *(Calcule p_i.)*

Paso 4 Si $|p - p_0| < TOL$ entonces

SALIDA (p); *(Procedimiento terminado satisfactoriamente).*

PARAR.

Paso 5 Tome $i = i + 1$.

Paso 6 Tome $p_0 = p$. (*Defina de nuevo* p_0.)

Paso 7 SALIDA ('El método fracasó después de N_0 iteraciones N_0 =', N_0);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

El siguiente ejemplo ilustra la técnica de la iteración funcional.

EJEMPLO 3 La ecuación $x^3 + 4x^2 - 10 = 0$ tiene una raíz única en [1, 2]. Hay muchas formas para convertirla en la forma $x = g(x)$ mediante un simple manejo algebraico. Por ejemplo, para obtener la función g que se describe en (c), podemos manejar la ecuación $x^3 + 4x^2 - 10 = 0$ así:

$$4x^2 = 10 - x^3, \quad \text{así} \quad x^2 = \frac{1}{4}(10 - x^3),$$

y

$$x = \pm\frac{1}{2}(10 - x^3)^{1/2}.$$

Para obtener una solución positiva, elegimos $g_3(x)$ como se muestra aquí. No es importante derivar las funciones que se indican, pero debemos verificar que el punto fijo de cada una sea realmente una solución de la ecuación original $x^3 + 4x^2 - 10 = 0$.

a. $x = g_1(x) = x - x^3 - 4x^2 + 10$

b. $x = g_2(x) = \left(\frac{10}{x} - 4x\right)^{1/2}$

c. $x = g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$

d. $x = g_4(x) = \left(\frac{10}{4 + x}\right)^{1/2}$

e. $x = g_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$

Con $p_0 = 1.5$, la tabla 2.2 proporciona los resultados del método de iteración de punto fijo para las cinco opciones de g.

La raíz real es 1.365230013, según se señaló en el ejemplo 1 de la sección 2.1. Al comparar los resultados del algoritmo de bisección que vienen en el ejemplo, observamos que se obtuvieron excelentes resultados con las opciones (c), (d) y (e), ya que el método de bisección requiere 27 iteraciones para garantizar la exactitud. Conviene señalar que la opción (a) ocasiona divergencia y que (b) se torna indefinida porque contiene la raíz cuadrada de un número negativo. ■

Aun cuando las funciones de este ejemplo son problemas de punto fijo para el mismo problema de búsqueda de raíz, difieren ampliamente como métodos para aproximar la so-

Tabla 2.2

n	(a)	(b)	(c)	(d)	(e)
0	1.5	1.5	1.5	1.5	1.5
1	−0.875	0.8165	1.286953768	1.348399725	1.373333333
2	6.732	2.9969	1.402540804	1.367376372	1.365262015
3	−469.7	$(-8.65)^{1/2}$	1.345458374	1.364957015	1.365230014
4	1.03×10^8		1.375170253	1.365264748	1.365230013
5			1.360094193	1.365225594	
6			1.367846968	1.365230576	
7			1.363887004	1.365229942	
8			1.365916734	1.365230022	
9			1.364878217	1.365230012	
10			1.365410062	1.365230014	
15			1.365223680	1.365230013	
20			1.365230236		
25			1.365230006		
30			1.365230013		

lución a este tipo de problemas. Su propósito es ilustrar la pregunta que es preciso contestar:

> ¿Cómo podemos encontrar un problema de punto fijo capaz de producir una sucesión que converga confiable y rápidamente en una solución en un problema de búsqueda de raíz?

El resultado siguiente y su corolario nos dan algunas pistas sobre los procedimientos que deberíamos seguir y, quizá lo más importante, algunos que debemos excluir.

Teorema 2.3 **(Teorema de punto fijo)**

Sea $g \in C[a, b]$ tal que $g(x) \in [a, b]$ para toda x en $[a, b]$. Además supongamos que existe g' en (a, b) y una constante positiva $0 < k < 1$ tales que

$$|g'(x)| \leq k, \qquad \text{para toda } x \in (a, b).$$

Entonces, para cualquier número p_0 en $[a, b]$, la sucesión definida por

$$p_n = g(p_{n-1}), \qquad n \geq 1,$$

converge al único punto fijo p en $[a, b]$. ■

Demostración El teorema 2.2 implica que existe un punto fijo único en $[a, b]$. Puesto que g transforma $[a, b]$ en sí mismo, la sucesión $\{p_n\}_{n=0}^{\infty}$ se define para toda $n \geq 0$ y $p_n \in [a, b]$ para toda n. Aplicando el hecho de que $|g'(x)| \leq k$ y el teorema del valor medio, tenemos, para cada n,

$$|p_n - p| = |g(p_{n-1}) - g(p)| = |g'(\xi_n)|\,|p_{n-1} - p| \leq k|p_{n-1} - p|,$$

donde $\xi_n \in (a, b)$. Al aplicar esta desigualdad inductivamente se obtiene

$$|p_n - p| \leq k|p_{n-1} - p| \leq k^2\,|p_{n-2} - p| \leq \ldots \leq k^n|p_0 - p|. \tag{2.4}$$

Puesto que $0 < k < 1$, se tiene $\lim_{n\to\infty} k^n = 0$ y

$$\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n |p_0 - p| = 0.$$

y $\{p_n\}_{n=0}^{\infty}$ converge a p. ■ ■ ■

Corolario 2.4 Si g satisface las hipótesis del teorema 2.3, las cotas del error que supone utilizar p_n para aproximar p están dadas por

$$|p_n - p| \leq k^n \text{ máx } \{p_0 - a, b - p_0\}$$

y por

$$|p_n - p| \leq \frac{k^n}{1-k} |p_1 - p_0|, \quad \text{para toda} \quad n \geq 1. \quad ■$$

Demostración Puesto que $p \in [a, b]$, la primera cota proviene de la desigualdad (2.4):

$$|p_n - p| \leq k^n |p_0 - p| \leq k^n \text{ máx } \{p_0 - a, b - p_0\}.$$

Para $n \geq 1$ el procedimiento empleado en la demostración del teorema 2.3 implica que

$$|p_{n+1} - p_n| = |g(p_n) - g(p_{n-1})| \leq k|p_n - p_{n-1}| \leq \cdots \leq k^n |p_1 - p_0|.$$

Por tanto, cuando $m > n \geq 1$,

$$\begin{aligned}|p_m - p_n| &= |p_m - p_{m-1} + p_{m-1} - \cdots + p_{n+1} - p_n| \\ &\leq |p_m - p_{m-1}| + |p_{m-1} - p_{m-2}| + \cdots + |p_{n+1} - p_n| \\ &\leq k^{m-1}|p_1 - p_0| + k^{m-2}|p_1 - p_0| + \cdots + k^n|p_1 - p_0| \\ &= k^n |p_1 - p_0|(1 + k + k^2 + \cdots + k^{m-n-1}).\end{aligned}$$

Por el teorema 2.3, $\lim_{m\to\infty} p_m = p$, de modo que

$$|p - p_n| = \lim_{m\to\infty} |p_m - p_n| \leq \lim_{m\to\infty} k^n |p_1 - p_0| \sum_{i=0}^{m-n-1} k^i \leq k^n |p_1 - p_0| \sum_{i=0}^{\infty} k^i.$$

Pero $\sum_{i=0}^{\infty} k^i$ es una serie geométrica con razón k y $0 < k < 1$. Esta sucesión converge a $1/(1 - k)$, lo que nos da la segunda cota:

$$|p - p_n| \leq \frac{k^n}{1-k} |p_1 - p_0|.$$

■ ■ ■

Ambas desigualdades del corolario relacionan la razón a la que $\{p_n\}_{n=0}^{\infty}$ converge a la cota k de la primera derivada. La razón de convergencia depende del factor k^n. Cuando más pequeño sea el valor de k, más rápida será la convergencia, la cual puede ser muy lenta si k está cerca de 1. En el siguiente ejemplo, reconsideraremos los métodos de punto fijo del ejemplo 3 a la luz de los resultados que describimos en el teorema 2.3 y su corolario.

EJEMPLO 4 **a.** Para $g_1(x) = x - x^3 - 4x^2 + 10$, tenemos $g_1(1) = 6$ y $g_1(2) = -12$, de modo que g_1 no mapea [1, 2] en sí mismo. Además, $g'_1(x) = 1 - 3x^2 - 8x$, de modo que $|g'_1(x)| > 1$ para toda x en [1, 2]. Aunque el teorema 2.3 no garantiza que el método deba fallar para esta elección de g, tampoco tenemos razón para esperar una convergencia.

b. Con $g_2(x) = [(10/x) - 4x]^{1/2}$, podemos ver que g_2 no mapea [1, 2] en [1, 2] y que la sucesión $\{p_n\}_{n=0}^{\infty}$ no está definida en $p_0 = 1.5$. Además, tampoco hay un intervalo que contenga a $p \approx 1.365$ tal que

$$|g_2'(x)| < 1, \qquad \text{puesto que} \qquad |g_2'(p)| \approx 3.4.$$

No hay razón para esperar que este método converja.

c. Para la función $g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$,

$$g_3'(x) = -\frac{3}{4}x^2(10 - x^3)^{-1/2} < 0 \qquad \text{en } [1, 2],$$

así que g_3 es estrictamente decreciente en [1, 2]. Sin embargo, $|g_3'(2)| \approx 2.12$, por lo cual la condición $|g_3'(x)| \leq k < 1$ falla en [1, 2]. Un análisis más cerca de la sucesión $\{p_n\}_{n=0}^{\infty}$, comenzando con $p_0 = 1.5$ revela que basta considerar el intervalo [1, 1.5] en vez de [1, 2]. En este intervalo sigue siendo verdad que $g_3'(x) < 0$ y g_3 es estrictamente decreciente, pero además

$$1 < 1.28 \approx g_3(1.5) \leq g_3(x) \leq g_3(1) = 1.5,$$

para toda $x \in [1, 1.5]$. Esto demuestra que g_3 mapea el intervalo [1, 1.5] en sí mismo. Puesto que también es cierto que $|g_3'(x)| \leq |g_3'(1.5)| \approx 0.66$ en este intervalo, el teorema 2.3 confirma la convergencia de la cual ya estábamos conscientes.

d. Para $g_4(x) = (10/(4 + x))^{1/2}$ tenemos

$$|g_4'(x)| = \left|\frac{-5}{\sqrt{10}(4 + x)^{3/2}}\right| \leq \frac{5}{\sqrt{10}(5)^{3/2}} < 0.15, \qquad \text{para toda} \quad x \in [1, 2].$$

La cota en la magnitud de $g_4'(x)$ es mucho menor que la de la magnitud de $g_3'(x)$ lo cual explica la convergencia más rápida que se obtiene con g_4.

e. La sucesión definida por

$$g_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$$

converge mucho más rápido que nuestras otras opciones. En las siguientes secciones veremos de dónde provino esta opción y por qué es tan efectiva. ■

CONJUNTO DE EJERCICIOS 2.2

1. Use el manejo algebraico para demostrar que las siguientes funciones tienen un punto fijo en p exactamente cuando $f(p) = 0$, donde $f(x) = x^4 + 2x^2 - x - 3$.

a. $g_1(x) = (3 + x - 2x^2)^{1/4}$ **b.** $g_2(x) = \left(\frac{x + 3 - x^4}{2}\right)^{1/2}$

c. $g_3(x) = \left(\dfrac{x+3}{x^2+2}\right)^{1/2}$ **d.** $g_4(x) = \dfrac{3x^4+2x^2+3}{4x^3+4x-1}$

2. **a.** Efectúe cuatro iteraciones, si es posible hacerlo, en las funciones g definidas en el ejercicio 1. Sea $p_0 = 1$ y $p_{n+1} = g(p_n)$ para $n = 0, 1, 2, 3$.

b. ¿Cuál función, a su juicio, dará la mejor aproximación a la solución?

3. Se proponen los tres métodos siguientes para calcular $21^{1/3}$. Clasifíquelos por orden, basándose para ello en la rapidez de convergencia y suponiendo que $p_0 = 1$.

a. $p_n = \dfrac{20p_{n-1} + 21/p_{n-1}^2}{21}$ **b.** $p_n = p_{n-1} - \dfrac{p_{n-1}^3 - 21}{3p_{n-1}^2}$

c. $p_n = p_{n-1} - \dfrac{p_{n-1}^4 - 21p_{n-1}}{p_{n-1}^2 - 21}$ **d.** $p_n = \left(\dfrac{21}{p_{n-1}}\right)^{1/2}$

4. Los cuatro siguientes métodos tienen por objeto calcular $7^{1/5}$. Clasifíquelos por orden, tomando como base su rapidez de convergencia y suponiendo que $p_0 = 1$.

a. $p_n = \left(1 + \dfrac{7 - p_{n-1}^3}{p_{n-1}^2}\right)^{1/2}$ **b.** $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{p_{n-1}^2}$

c. $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{5p_{n-1}^4}$ **d.** $p_n = p_{n-1} - \dfrac{p_{n-1}^5 - 7}{12}$

5. Aplique el método de iteración de punto fijo para determinar una solución con una exactitud de 10^{-2} para $x^4 - 3x^2 - 3 = 0$ en $[1, 2]$. Utilice $p_0 = 1$.

6. Aplique un método de iteración de punto fijo para determinar una solución exacta dentro de 10^{-2} para $x^3 - x - 1 = 0$ en $[1, 2]$. Utilice $p_0 = 1$.

7. Aplique el teorema 2.2 para demostrar que $g(x) = \pi + 0.5\,\text{sen}(x/2)$ tiene un único punto fijo en $[0, 2\pi]$. Use la iteración de punto fijo para obtener una aproximación al punto fijo con una exactitud de 10^{-2}. Use el corolario 2.4 para estimar la cantidad de iteraciones necesarias para lograr una exactitud de 10^{-4} y después compare esta estimación teórica con la cantidad que realmente se requiere.

8. Aplique el teorema 2.2 para demostrar que $g(x) = 2^{-x}$ tiene un único punto fijo en $[\frac{1}{3}, 1]$. Utilice la iteración de punto fijo para obtener una aproximación del punto fijo exacta en 10^{-4}. Use el corolario 2.4 para estimar la cantidad de iteraciones necesarias para alcanzar una exactitud de 10^{-4} y después compare esta estimación teórica con la cantidad que realmente se requiere.

9. Aplique un método de iteración de punto fijo para obtener una aproximación a $\sqrt{3}$ con una exactitud de 10^{-4}. Compare su resultado con el número de iteraciones que requiere la respuesta obtenida en el ejercicio 10 de la sección 2.1.

10. Use un método de iteración de punto fijo para obtener una aproximación a $\sqrt[3]{25}$ con una exactitud de 10^{-4}. Compare su resultado con el número de iteraciones que requiere la respuesta obtenida en el ejercicio 11 de la sección 2.1.

11. En cada una de las siguientes ecuaciones, determine un intervalo $[a, b]$ en que convergerá la iteración de punto fijo. Estime la cantidad de iteraciones necesarias para obtener aproximaciones con una exactitud de 10^{-5} y realice los cálculos.

a. $x = \dfrac{2 - e^x + x^2}{3}$ **b.** $x = \dfrac{5}{x^2} + 2$

c. $x = (e^x/3)^{1/2}$ **d.** $x = 5^{-x}$

e. $x = 6^{-x}$ **f.** $x = 0.5(\text{sen}\,x + \cos x)$

12. En cada una de las siguientes ecuaciones, determine una función g y un intervalo $[a, b]$ donde la iteración de punto fijo convergerá en una solución positiva de la ecuación.

a. $3x^2 - e^x = 0$ **b.** $x - \cos x = 0$

Obtenga las soluciones con una exactitud de 10^{-5}.

13. Encuentre todos los ceros de $f(x) = x^2 + 10\cos x$ aplicando el método de iteración de punto fijo para una función de iteración apropiada g. Encuentre los ceros con una exactitud de 10^{-4}.

14. Aplique el método de iteración de punto fijo para determinar una solución con una exactitud de 10^{-4} con $x = \tan x$, para x en $[4, 5]$.

15. Aplique el método de iteración de punto fijo para determinar una solución con una exactitud de 10^{-2} para $2 \operatorname{sen} \pi x + x = 0$ en $[1, 2]$. Use $p_0 = 1$.

16. Sea A una constante positiva y $g(x) = 2x - Ax^2$.

a. Demuestre que si una iteración de punto fijo converge a un límite diferente de cero, entonces el límite es $p = 1/A$, de modo que la inversa de un número puede obtenerse usando sólo multiplicaciones y sustracciones.

b. Encuentre un intervalo alrededor de $1/A$ donde converja una iteración de punto fijo, a condición de que p_0 se encuentre en ese intervalo.

17. Encuentre una función g definida en $[0, 1]$ que no satisfaga ninguna de las hipótesis del teorema 2.2, pero que siga teniendo un punto fijo único en $[0, 1]$.

18. **a.** Demuestre que el teorema 2.2 es verdadero si la desigualdad $|g'(x)| \le k$ se reemplaza con $g'(x) \le k$ para toda $x \in (a, b)$. [*Sugerencia*: sólo se pone en tela de juicio la unicidad].

b. Demuestre que el teorema 2.3 no es válido si la desigualdad $|g'(x)| \le k$ se reemplaza con $g'(x) \le k$. [*Sugerencia*: demuestre que $g(x) = 1 - x^2$ para x en $[0, 1]$, proporciona un contra- ejemplo].

19. **a.** Aplique el teorema 2.3 para demostrar que la sucesión definida por

$$x_n = \frac{1}{2}x_{n-1} + \frac{1}{x_{n-1}}, \qquad \text{para} \quad n \ge 1,$$

converge a $\sqrt{2}$ siempre que $x_0 > \sqrt{2}$.

b. Aplique el hecho de que $0 < (x_0 - \sqrt{2})^2$ siempre que $x_0 \neq \sqrt{2}$ para demostrar que si $0 < x_0 < \sqrt{2}$, entonces $x_1 > \sqrt{2}$.

c. Utilice los resultados de las partes (a) y (b) para demostrar que la sucesión en (a) converge a $\sqrt{2}$ siempre que $x_0 > 0$.

20. **a.** Demuestre que si A es un número positivo, entonces la sucesión definida por medio de

$$x_n = \frac{1}{2}x_{n-1} + \frac{A}{2x_{n-1}}, \qquad \text{para} \quad n \ge 1,$$

converge a $\sqrt{A}$ siempre que $x_0 > 0$.

b. ¿Qué sucede si $x_0 < 0$?

21. Reemplace la suposición del teorema 2.3 de que "existe un número positivo $k < 1$ con $|g'(x)| \le k$" con "g satisface la condición de Lipschitz en el intervalo $[a, b]$ con la constante de Lipschitz $L < 1$". (Véase el ejercicio 25, sección 1.1.) Demuestre que las conclusiones de este teorema siguen siendo válidas.

22. Suponga que g es continuamente diferenciable en algún intervalo (c, d) que contenga el punto fijo p de g. Demuestre que si $|g'(p)| < 1$, entonces existe una $\delta > 0$ tal que la iteración de punto fijo converge para cualquier aproximación p_0 siempre que $|p_0 - p| \le \delta$.

23. Un objeto que cae verticalmente en el aire está sujeto a una resistencia viscosa y también a la fuerza de gravedad. Suponga que dejamos caer un objeto de masa m desde una altura s_0 y que la altura del objeto después de t segundos es

$$s(t) = s_0 - \frac{mg}{k}t + \frac{m^2g}{k^2}(1 - e^{-kt/m}),$$

donde $g = 32.17$ pies/s^2 y k representa el coeficiente de resistencia del aire en lb-s/pies. Suponga que $s_0 = 300$ pies, $m = 0.25$ lb, y que $k = 0.1$ lb-s/pies. Calcule, con una exactitud de 0.01 s, el tiempo que tarda este peso de un cuarto de libra en caer al suelo.

24. Sea $g \in C^1[a, b]$ y supongamos que p está en (a, b) con $g(p) = p$ y que $|g'(p)| > 1$. Demuestre que existe una $\delta > 0$ tal que, si $0 < |p_0 - p| < \delta$, entonces $|p_0 - p| < |p_1 - p|$. Por tanto, sin importar qué tan cerca esté la aproximación inicial p_0 de p, la siguiente iteración p_1 se encontrará más lejos, de modo que la iteración de punto fijo no converge si $p_0 \neq p$.

2.3 El método de Newton

El **método de Newton** (o **método de Newton-Raphson**) es una de las técnicas numéricas para resolver un problema de búsqueda de raíces $f(x) = 0$ más poderosas y conocidas. Hay muchas formas de introducirlo. La más común consiste en considerarlo gráficamente. Otra posibilidad consiste en derivarlo como una técnica que permite lograr una convergencia más rápida que la que ofrecen otros tipos de iteración funcional. Esto lo hacemos en la sección 2.4. Una tercera forma de introducir el método de Newton, que estudiaremos a continuación, se basa en los polinomios de Taylor.

Supongamos que $f \in C^2[a, b]$. Sea $\bar{x} \in [a, b]$ una aproximación de p tal que $f'(\bar{x}) \neq 0$ y $|p - \bar{x}|$ es "pequeño". Consideremos el primer polinomio de Taylor para $f(x)$ expandido alrededor de $\bar{x}$,

$$f(x) = f(\bar{x}) + (x - \bar{x})f'(\bar{x}) + \frac{(x - \bar{x})^2}{2} f''(\xi(x)),$$

donde $\xi(x)$ está entre x y $\bar{x}$. Dado que $f(p) = 0$ esta ecuación, con $x = p$, da

$$0 = f(\bar{x}) + (p - \bar{x})f'(\bar{x}) + \frac{(p - \bar{x})^2}{2} f''(\xi(p)).$$

Derivamos el método de Newton suponiendo que, como $|p - \bar{x}|$ es tan pequeño, el término que contiene $(p - \bar{x})^2$ es mucho menor y que

$$0 \approx f(\bar{x}) + (p - \bar{x})f'(\bar{x}).$$

Despejando p de esta ecuación obtenemos

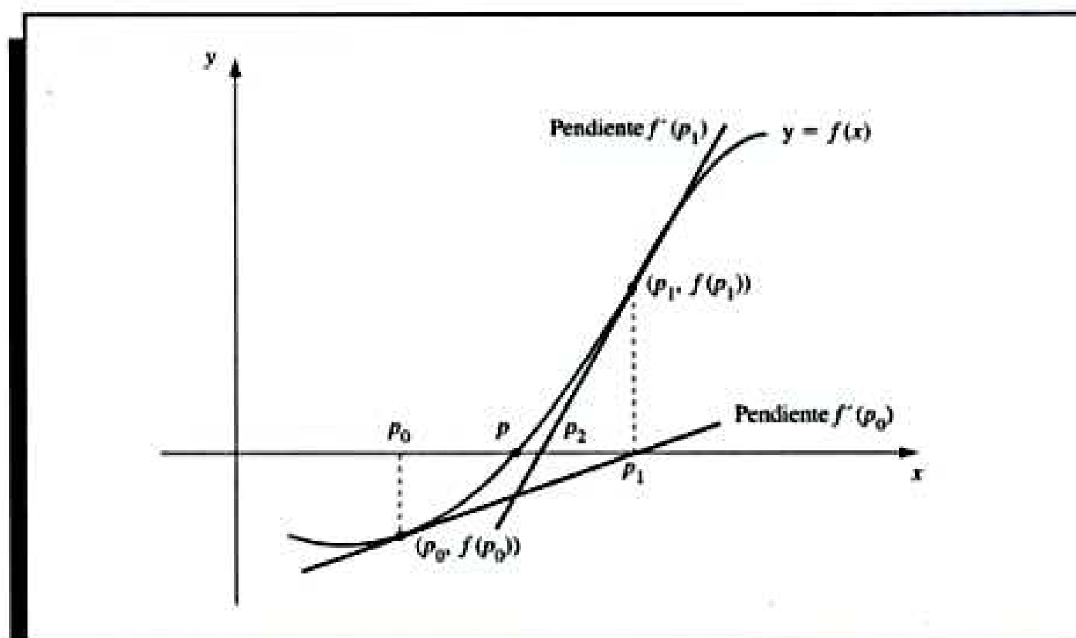
$$p \approx \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}.$$

Esto nos prepara para introducir el método de Newton, el cual comienza con una aproximación inicial p_0 y genera la sucesión $\{p_n\}_{n=0}^{\infty}$ definida por

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \qquad \text{para} \quad n \geq 1. \tag{2.5}$$

La figura 2.7 muestra gráficamente cómo se obtienen las aproximaciones usando tangentes sucesivas. (Véase también el ejercicio 11.) Comenzando con la aproximación inicial p_0, la aproximación p_1 es la intersección con el eje x de la línea tangente a la gráfica de f en $(p_0, f(p_0))$. La aproximación p_2 es la intersección x de la tangente a la gráfica de f en $(p_1, f(p_1))$ y así sucesivamente. El algoritmo 2.3 sigue este procedimiento.

Figura 2.7



ALGORITMO 2.3

Método de Newton

Para obtener una solución a $f(x) = 0$ dada la función diferenciable f y una aproximación inicial p_0:

ENTRADA aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de fracaso.

Paso 1 Tome $i = 1$.

Paso 2 Mientras $i \leq N_0$ haga pasos 3–6.

Paso 3 Tome $p = p_0 - f(p_0)/f'(p_0)$. (Calcule p_i.)

Paso 4 Si $|p - p_0| < TOL$ entonces
SALIDA (p); (*Procedimiento terminado satisfactoriamente.*)
PARAR.

Paso 5 Tome $i = i + 1$.

Paso 6 Tome $p_0 = p$. (*Redefina p_0.*)

Paso 7 SALIDA ('El método fracasó después de N_0 iteraciones, N_0 = ', N_0);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Las desigualdades de la técnica de paro dadas con el método de bisección son aplicables al método de Newton. Es decir, seleccione una tolerancia $\varepsilon > 0$ y construya $p_1, \ldots p_N$ hasta que

$$|p_N - p_{N-1}| < \varepsilon, \tag{2.6}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \varepsilon, \quad p_N \neq 0, \tag{2.7}$$

o bien

$$|f(p_N)| < \varepsilon. \tag{2.8}$$

Una forma de la desigualdad (2.6) se usa en el paso 4 del algoritmo 2.3. Obsérvese que tal vez la desigualdad (2.8) no proporcione mucha información sobre el error real $|p_N - p|$. (Consúltese el ejercicio 14 en la sección 2.1.)

El método de Newton es una técnica de iteración funcional de la forma $p_n = g(p_{n-1})$, para la cual

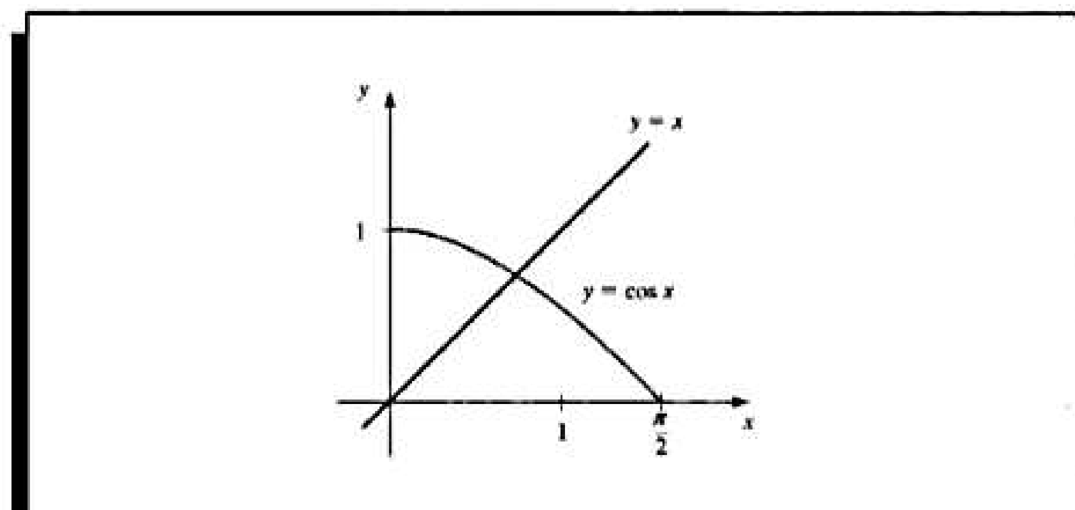
$$g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \qquad \text{para } n \geq 1. \tag{2.9}$$

De hecho, ésta es la técnica de iteración funcional que usamos para obtener la rápida convergencia que vimos en la parte (e) del ejemplo 3 en la sección 2.2.

De la ecuación (2.9) inferimos que no es posible continuar con el método de Newton si $f'(p_{n-1}) = 0$ para alguna n. De hecho, veremos que el método es más efectivo cuando f' está acotada lejos de cero y cerca de p.

EJEMPLO 1 Suponga que queremos aproximar un punto fijo de $g(x) = \cos x$. La gráfica de la figura 2.8 implica que sólo hay un punto fijo en $[0, \pi/2]$.

Figura 2.8



La tabla 2.3 muestra los resultados de la iteración de punto fijo con $p_0 = \pi/4$. Lo más que podemos concluir de estos resultados es que $p \approx 0.74$.

Para otro enfoque de este problema, definamos $f(x) = \cos x - x$ y apliquemos el método de Newton. Como $f'(x) = -\text{sen}\, x - 1$, la sucesión es generada por

$$p_n = p_{n-1} - \frac{\cos p_{n-1} - p_{n-1}}{-\text{sen}\, p_{n-1} - 1}, \qquad \text{para } n \geq 1.$$

Para $p_0 = \pi/4$, se generan las aproximaciones de la tabla 2.4. Se obtiene una excelente aproximación con $n = 3$. Sería de esperar que este resultado tenga una precisión del número de cifras enumeradas, debido a la coincidencia de p_3 y p_4. ■

Tabla 2.3

n	p_n
0	0.7853981635
1	0.7071067810
2	0.7602445972
3	0.7246674808
4	0.7487198858
5	0.7325608446
6	0.7434642113
7	0.7361282565

Tabla 2.4

n	p_n
0	0.7853981635
1	0.7395361337
2	0.7390851781
3	0.7390851332
4	0.7390851332

La derivación del método de Newton por medio de las series de Taylor, tema que tratamos al inicio de esta sección, subraya la importancia de una aproximación inicial exacta. La suposición fundamental es que el término que contiene $(p - \bar{x})^2$ es, en comparación con $|p - \bar{x}|$, tan pequeño que podemos suprimirlo. Esto evidentemente sería falso a menos que $\bar{x}$ sea una buena aproximación de p. En particular, si p_0 no se acerca bastante a la raíz real, el método de Newton quizá no converja a la raíz. Pero no siempre es así. (En los ejercicios 12 y 16 se ejemplifican algunas de las posibilidades que pueden presentarse.)

El siguiente teorema de convergencia para el método de Newton muestra la importancia teórica de la elección de p_0.

Teorema 2.5 Sea $f \in C^2[a, b]$. Si $p \in [a, b]$ es tal que $f(p) = 0$ y $f'(p) \neq 0$, entonces existe $\delta > 0$ tal que el método de Newton genera una sucesión $\{p_n\}_{n=1}^{\infty}$ que converge a p para cualquier aproximación inicial $p_0 \in [p - \delta, p + \delta]$. ■

Demostración La demostración se basa en analizar el método de Newton como un esquema de iteración funcional $p_n = g(p_{n-1})$, para $n \geq 1$, con

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Sea k un número cualquiera en $(0, 1)$. Primero debemos encontrar un intervalo $[p - \delta, p + \delta]$ que g mapea en sí mismo y en que $|g'(x)| \leq k$ para toda $x \in (p - \delta, p + \delta)$.

Como $f'(p) \neq 0$ y f' es continua, existe $\delta_1 > 0$ tal que $f'(x) \neq 0$ para $x \in [p - \delta_1, p + \delta_1] \subseteq [a, b]$. Por tanto, g está definida y es continua en $[p - \delta_1, p + \delta_1]$. También,

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2},$$

para $x \in [p - \delta_1, p + \delta_1]$ y como $f \in C^2[a, b]$, tendremos $g \in C^1[p - \delta_1, p + \delta_1]$.

Por suposición, $f(p) = 0$, así que

$$g'(p) = \frac{f(p)f''(p)}{[f'(p)]^2} = 0.$$

Como g' es continua y como $0 < k < 1$, parte (b) del ejercicio 27 en la sección 1.1 implica que existe δ, con $0 < \delta < \delta_1$, y

$$|g'(x)| \leq k, \quad \text{para toda } x \in [p - \delta, p + \delta].$$

Todavía falta demostrar que g manda $[p - \delta, p + \delta]$ en $[p - \delta, p + \delta]$. Si $x \in [p - \delta, p + \delta]$, el teorema del valor medio implica que, para algún número ξ entre x y p, $|g(x) - g(p)| = |g'(\xi)||x - p|$. Por tanto,

$$|g(x) - p| = |g(x) - g(p)| = |g'(\xi)||x - p| \leq k|x - p| < |x - p|.$$

Puesto que $x \in [p - \delta, p + \delta]$, se deduce que $|x - p| < \delta$ y que $|g(x) - p| < \delta$. Este resultado implica que g manda $[p - \delta, p + \delta]$ en $[p - \delta, p + \delta]$.

Todas las hipótesis del teorema del punto fijo se satisfacen, de modo que la sucesión $\{p_n\}_{n=1}^{\infty}$ definida por

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \qquad \text{para} \quad n \geq 1,$$

converge a p para cualquier $p_0 \in [p - \delta, p + \delta]$. ■ ■ ■

El teorema 2.5 establece que, bajo suposiciones razonables, el método de Newton converge a condición de que se escoja una aproximación inicial suficientemente exacta. También implica que la constante k que acota la derivada de g y, en consecuencia, indica la rapidez de convergencia del método, disminuye a cero a medida que el procedimiento avanza. Este resultado es importante para la teoría del método de Newton, aunque pocas veces se aplica en la práctica, ya que no nos dice cómo determinar δ. En una aplicación práctica, se elige una aproximación inicial y las aproximaciones sucesivas se generan mediante el método de Newton. Por lo general, esto convergerá rápidamente a la raíz, o será claro que la convergencia es improbable.

El método de Newton es una técnica muy poderosa, pero presenta un grave problema: la necesidad de conocer el valor de la derivada de f en cada aproximación. Con frecuencia es más difícil determinar $f'(x)$ y se requieren más operaciones aritméticas para calcularlo que para $f(x)$.

Si queremos evitar el problema de evaluar la derivada en el método de Newton, derivamos una pequeña variación. Por definición,

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}}.$$

Haciendo $x = p_{n-2}$, tenemos

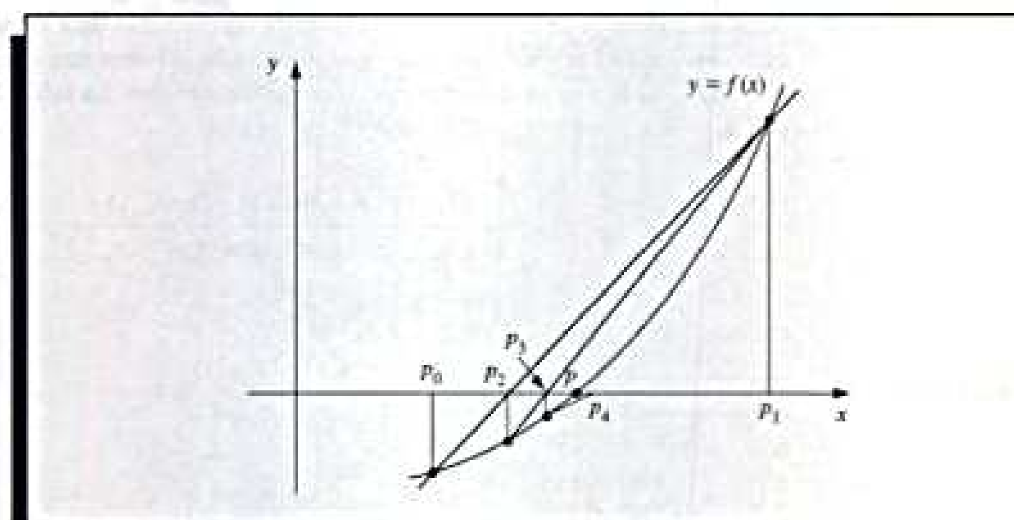
$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}} = \frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}.$$

Al aplicar esta aproximación para $f'(p_{n-1})$ en la fórmula de Newton, se obtiene

$$p_n = p_{n-1} - \frac{f(p_{n-1})(p_{n-1} - p_{n-2})}{f(p_{n-1}) - f(p_{n-2})}. \tag{2.10}$$

La técnica que utiliza esta fórmula recibe el nombre de **método de la secante** y se incluye en el algoritmo 2.4 (véase Fig. 2.9). Comenzando con las dos aproximaciones iniciales p_0 y p_1, la aproximación p_2 es la intersección del eje x y la línea que une $(p_0, f(p_0))$ y $(p_1, f(p_1))$. La aproximación p_3 es la intersección del eje x y la línea que une $(p_1, f(p_1))$ y $(p_2, f(p_2))$, y así sucesivamente.

Figura 2.9



ALGORITMO 2.4

Secante

Para encontrar una solución para $f(x) = 0$ dadas las aproximaciones iniciales p_0 y p_1:

ENTRADA aproximaciones iniciales p_0, p_1; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de fracaso.

Paso 1 Tome $i = 2$;
$q_0 = f(p_0)$;
$q_1 = f(p_1)$.

Paso 2 Mientras $i \leq N_0$ haga pasos 3-6.

Paso 3 Tome $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$. (*Calcule* p_i.)

Paso 4 Si $|p - p_1| < TOL$ entonces
SALIDA (p); (*Procedimiento terminado satisfactoriamente.*)
PARAR.

Paso 5 Tome $i = i + 1$.

Paso 6 Tome $p_0 = p_1$; (*Redefina* p_0, q_0, p_1, q_1.)
$q_0 = q_1$;
$p_1 = p$;
$q_1 = f(p)$.

Paso 7 SALIDA ('El método falló después de N_0 iteraciones, N_0 = ', N_0);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

El siguiente ejemplo incluye un problema que vimos en el ejemplo 1, donde aplicamos el método de Newton con $p_0 = \pi/4$.

EJEMPLO 2 Aplique el método de la secante para encontrar una solución de $x = \cos x$. En el ejemplo 1 comparamos la iteración funcional y el método de Newton con la aproximación inicial $p_0 = \pi/4$. Aquí necesitamos dos aproximaciones iniciales. La tabla 2.5 lista los cálculos con $p_0 = 0.5$, $p_1 = \pi/4$ y la fórmula

$$p_n = p_{n-1} - \frac{(p_{n-1} - p_{n-2})(\cos p_{n-1} - p_{n-1})}{(\cos p_{n-1} - p_{n-1}) - (\cos p_{n-2} - p_{n-2})}, \qquad \text{para} \quad n \geq 2,$$

tomada del algoritmo 2.4. ■

Tabla 2.5

n	p_n
0	0.5
1	0.7853981635
2	0.7363841388
3	0.7390581392
4	0.7390851493
5	0.7390851332

Al comparar los resultados de ahora con los del ejemplo 1 observamos que p_5 es exacto hasta la décima cifra decimal. Nótese que la convergencia del método de la secante es más rápida que la iteración funcional, pero un poco más lenta en este ejemplo que el método de Newton, en el cual obtuvimos este grado de exactitud con p_3. Este resultado generalmente es verdadero. (Ver el ejercicio 12 de la sección 2.4.)

El método de Newton o el método de la secante a menudo se usan para refinar las respuestas conseguidas con otra técnica, como el método de bisección. Dado que el método de Newton requiere de una buena aproximación inicial, pero por lo general da una convergencia rápida, sirve perfectamente para el propósito antes mencionado.

Cada par sucesivo de aproximaciones en el método de bisección encierra una raíz p de la ecuación; es decir, para cada entero positivo n, una raíz se encuentra entre a_n y b_n. Ello significa que para cada n las iteraciones del método de bisección satisfacen

$$|p_n - p| < \frac{1}{2}|a_n - b_n|,$$

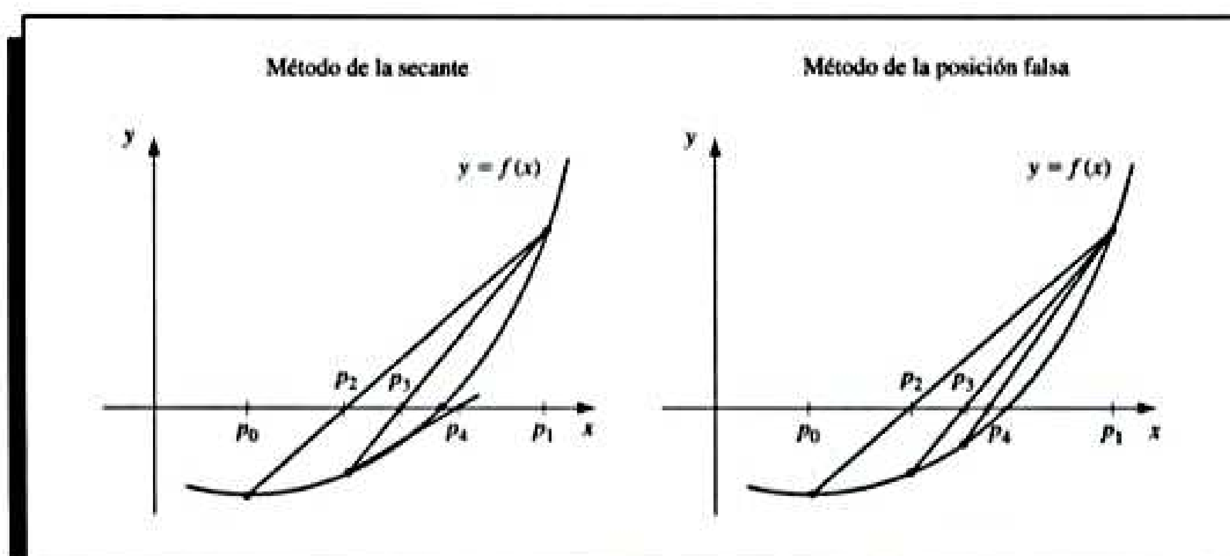
con lo cual se obtiene una cota de error para las aproximaciones fácilmente calculables. El acorralamiento de las raíces no está garantizado en el método de Newton ni en el de la secante. La tabla 2.4 contiene los resultados que se obtienen con el método de Newton aplicado a $f(x) = \cos x - x$, donde se comprobó que una raíz aproximada era 0.7390851332. Obsérvese que esta raíz no se encuentra entre p_0, p_1 o p_1, p_2. Las aproximaciones del método de la secante aplicadas a este problema vienen en la tabla 2.5. Las aproximaciones iniciales p_0 y p_1 se eligieron para encerrar la raíz, pero esto no lo hace el par de aproximaciones p_3 y p_4.

El **método de la posición falsa** (llamado también *método de Regla falsa*) genera aproximaciones del mismo modo que el de la secante, pero ofrece una prueba para asegurarse de que la raíz quede entre dos iteraciones sucesivas. Aunque no es un método que recomendamos, ilustra la forma en que puede incorporarse el acorralamiento.

Primero elegimos las aproximaciones iniciales p_0 y p_1 con $f(p_0) \cdot f(p_1) < 0$. La aproximación p_2 se escoge de la misma manera que en el método de la secante: como la intersección en x de la línea que une $(p_0, f(p_0))$ y $(p_1, f(p_1))$. Para decidir con cuál secante

calcularemos p_3 verificamos $f(p_2) \cdot f(p_1)$. Si este valor es negativo, entonces p_1 y p_2 encierran una raíz, y elegiremos p_3 como la intersección en x de la recta que une $(p_1, f(p_1))$ y $(p_2, f(p_2))$. Si no, elegimos p_3 como la intersección con x de la recta que une $(p_0, f(p_0))$ y $(p_2, f(p_2))$ y después intercambiamos los índices de p_0 y p_1. En forma análoga, una vez encontrado p_3, el signo de $f(p_3) \cdot f(p_2)$ determina si usaremos p_2 y p_3 o p_3 y p_1 para calcular p_4. En el segundo caso se reetiquetan p_2 y p_1. Con ello nos aseguramos de que la raíz quede entre iteraciones consecutivas. El proceso se describe en el algoritmo 2.5 y la figura 2.10 muestra cómo las iteraciones pueden diferir de las del método de la secante. En este ejemplo, las tres primeras aproximaciones son iguales, pero la cuarta es diferente.

Figura 2.10



ALGORITMO 2.5

Método de la posición falsa

Para encontrar una solución a $f(x) = 0$ dada la función continua f en el intervalo $[p_0, p_1]$ donde $f(p_0)$ y $f(p_1)$ tienen signos opuestos:

ENTRADA aproximaciones iniciales p_0, p_1 tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Tome $i = 2$;
$q_0 = f(p_0)$;
$q_1 = f(p_1)$.

Paso 2 Mientras $i \leq N_0$ haga pasos 3-7.

Paso 3 Tome $p = p_1 - q_1(p_1 - p_0)/(q_1 - q_0)$. (*Calcule* p_i.)

Paso 4 Si $|p - p_1| < TOL$ entonces
SALIDA (p); (*Procedimiento terminado satisfactoriamente.*)
PARAR.

Paso 5 Tome $i = i + 1$;
$q = f(p)$.

Paso 6 Si $q \cdot q_1 < 0$ entonces tome $p_0 = p_1$;
$q_0 = q_1$.

Paso 7 Tome $p_1 = p$;
$q_1 = q$.

Paso 8 SALIDA ('El método falló después de N_0 iteraciones, N_0 = ', N_0);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

EJEMPLO 3 La tabla 2.6 contiene los resultados del método de la posición falsa aplicado a $f(x) = \cos x - x$ con las mismas aproximaciones iniciales que utilizamos para el método de la secante en el ejemplo 2. Nótese que las aproximaciones son iguales hasta p_3 y que el método de la posición falsa requiere una iteración más para alcanzar la misma exactitud que el de la secante. ■

Tabla 2.6

n	p_n
0	0.5
1	0.7853981635
2	0.7363841388
3	0.7390581392
4	0.7390848638
5	0.7390851305
6	0.7390851332

La seguridad adicional que ofrece el método de la posición falsa suele requerir más cálculos que el método de la secante, del mismo modo que la simplificación que se logra con este último método respecto a aquél suele lograrse a costa de iteraciones adicionales. Al resolver los ejercicios 13 y 14, se verán más ejemplos de los aspectos positivos y negativos de ambos métodos.

CONJUNTO DE EJERCICIOS 2.3

1. Sean $f(x) = x^2 - 6$ y $p_0 = 1$. Aplique el método de Newton para encontrar p_2.
2. Sean $f(x) = -x^3 - \cos x$ y $p_0 = -1$. Aplique el método de Newton para encontrar p_2. ¿Podríamos utilizar $p_0 = 0$?
3. Sea $f(x) = x^2 - 6$. Con $p_0 = 3$ y $p_1 = 2$ encuentre p_3.
   a. Aplique el método de la secante.
   b. Aplique el método de la posición falsa.
   c. ¿Está (a) o (b) más cerca de $\sqrt{6}$?
4. Sea $f(x) = -x^3 - \cos x$. Con $p_0 = -1$ y $p_1 = 0$ obtenga p_3.
   a. Aplique el método de la secante.
   b. Aplique el método de la posición falsa.

5. Aplique el método de Newton para obtener soluciones con una exactitud de 10^{-4} para los siguientes problemas.

   a. $x^3 - 2x^2 - 5 = 0$, $[1, 4]$

   b. $x^3 + 3x^2 - 1 = 0$, $[-3, -2]$

   c. $x - \cos x = 0$, $[0, \pi/2]$

   d. $x - 0.8 - 0.2 \operatorname{sen} x = 0$, $[0, \pi/2]$

6. Aplique el método de Newton para obtener soluciones con una exactitud de 10^{-5} para los siguientes problemas:

   a. $e^x + 2^{-x} + 2\cos x - 6 = 0$ para $1 \le x \le 2$

   b. $\ln(x-1) + \cos(x-1) = 0$ para $1.3 \le x \le 2$

   c. $2x\cos 2x - (x-2)^2 = 0$ para $2 \le x \le 3$ y para $3 \le x \le 4$

   d. $(x-2)^2 - \ln x = 0$ para $1 \le x \le 2$ y para $e \le x \le 4$

   e. $e^x - 3x^2 = 0$ para $0 \le x \le 1$ y para $3 \le x \le 5$

   f. $\operatorname{sen} x - e^{-x} = 0$ para $0 \le x \le 1$, para $3 \le x \le 4$, y para $6 \le x \le 7$

7. Repita el ejercicio 5 usando (i) el método de la secante y (ii) el método de la posición falsa.

8. Repita el ejercicio 6 usando (i) el método de la secante y (ii) el método de la posición falsa.

9. Use el método de Newton para aproximar, con una exactitud de 10^{-4}, el valor de x que en la gráfica de $y = x^2$ produce el punto más cercano a $(1, 0)$ [*Sugerencia*: reduzca al mínimo $[d(x)]^2$ donde $d(x)$ representa la distancia entre (x, x^2) y $(1, 0)$.]

10. Con el método de Newton aproxime, con un grado de exactitud de 10^{-4}, el valor de x que en la gráfica de $y = 1/x$ produce el punto más cercano a $(2, 1)$.

11. Lo siguiente describe gráficamente el método de Newton: supongamos que existe $f'(x)$ y que no se anula en $[a, b]$. Supongamos, además, que existe una $p \in [a, b]$ tal que $f(p) = 0$, y sea $p_0 \in [a, b]$ arbitrario. Sea p_1 el punto donde la tangente a f en $(p_0, f(p_0))$ cruza el eje x. Para cada $n \ge 1$ sea p_n la intersección en x de la tangente a f en $(p_{n-1}, f(p_{n-1}))$. Derive la fórmula que describe este método.

12. Con el método de Newton resuelva la ecuación

$$0 = \frac{1}{2} + \frac{1}{4}x^2 - x \operatorname{sen} x - \frac{1}{2}\cos 2x, \qquad \text{con } p_0 = \frac{\pi}{2}.$$

Itere usando el método de Newton hasta lograr una exactitud de 10^{-5}. Explique por qué el resultado parece poco usual para el método de Newton. También resuelva la ecuación con $p_0 = 5\pi$ y $p_0 = 10\pi$.

13. El polinomio de cuarto grado

$$f(x) = 230x^4 + 18x^3 + 9x^2 - 221x - 9$$

tiene dos ceros reales, uno en $[-1, 0]$ y el otro en $[0, 1]$. Trate de aproximar estos ceros con una exactitud de 10^{-6} por medio de

   a. El método de la posición falsa

   b. El método de la secante

   c. El método de Newton

Utilice los extremos de cada intervalo como aproximaciones iniciales en (a) y en (b) y los intermedios como aproximaciones iniciales en (c).

14. La función $f(x) = \tan \pi x - 6$ tiene un cero en $(1/\pi)\arctan 6 \approx 0.447431543$. Sean $p_0 = 0$ y $p_1 = 0.48$, use diez iteraciones de cada uno de los siguientes métodos para aproximar esta raíz. ¿Cuál de ellos es más eficaz y por qué?

   a. Método de bisección

   b. Método de la posición falsa

   c. Método de la secante

15. La ecuación de iteración para el método de la secante puede escribirse en la forma más simple

$$p_n = \frac{f(p_{n-1})p_{n-2} - f(p_{n-2})p_{n-1}}{f(p_{n-1}) - f(p_{n-2})}.$$

Explique por qué, en términos generales, esta ecuación tiende a ser menos precisa que la del algoritmo 2.4.

16. La ecuación $x^2 - 10 \cos x = 0$ tiene dos soluciones ± 1.3793646. Con el método de Newton aproxime las soluciones con un grado de exactitud de 10^{-5} para los siguientes valores de p_0.

a. $p_0 = -100$ **c.** $p_0 = -25$ **e.** $p_0 = 50$

b. $p_0 = -50$ **d.** $p_0 = 25$ **f.** $p_0 = 100$

17. Use Maple para determinar cuántas iteraciones del método de Newton con $p_0 = \pi/4$ se necesitan para encontrar un cero de $f(x) = \cos x - x$ con una exactitud de 10^{-100}.

18. Repita el ejercicio 17 con $p_0 = \frac{1}{2}$, $p_1 = \frac{\pi}{4}$, y el método de la secante.

19. La función descrita por $f(x) = \ln(x^2 + 1) - e^{0.4x} \cos \pi x$ tiene una cantidad infinita de ceros.

a. Determine, con una exactitud de 10^{-6} el único cero negativo.

b. Determine, con una exactitud de 10^{-6} los cuatro ceros positivos más pequeños.

c. Determine una aproximación inicial razonable para obtener el n-ésimo cero positivo más pequeño de f [*Sugerencia*: trace una gráfica aproximada de f.]

d. Use la parte (c) para determinar, con una exactitud de 10^{-6} el 25º cero positivo más pequeño de f.

20. Encuentre una aproximación de λ con una exactitud de 10^{-4} para la ecuación de población

$$1\,564\,000 = 1\,000\,000e^{\lambda} + \frac{435\,000}{\lambda}(e^{\lambda} - 1),$$

que explicamos en la introducción del capítulo. Use este valor para predecir la población que habrá al final del segundo año, suponiendo que la tasa de inmigración durante este año se mantiene en 435 000 personas por año.

21. La suma de dos números es 20. Si cada uno se agrega a su raíz cuadrada, el producto de las dos sumas es 155.55. Determine los dos números con una exactitud de 10^{-4}.

22. El valor acumulado de una cuenta de ahorros que se basa en pagos periódicos puede calcularse con la *ecuación de anualidad* vencida

$$A = \frac{P}{i}[(1 + i)^n - 1].$$

En esta ecuación, A es el monto de la cuenta, P es la cantidad que se deposita periódicamente e i es la tasa de interés por periodo para los n periodos de depósito. A un ingeniero le gustaría tener una cuenta de ahorros con un monto de 750 000 dólares al momento de retirarse dentro de 20 años, y puede depositar 1 500 dólares mensuales para lograr dicho objetivo. ¿Cuál es la tasa mínima de interés a que puede invertirse ese dinero, suponiendo que es un interés compuesto mensual?

23. Los problemas relativos al dinero necesario para pagar una hipoteca de una casa durante un periodo fijo de tiempo requieren la fórmula

$$A = \frac{P}{i}[1 - (1 + i)^{-n}],$$

denominada *ecuación de la anualidad ordinaria*. En esta ecuación, A es el importe de la hipoteca, P es el importe de cada pago e i es la tasa de interés por periodo para n periodos. Supongamos que se necesita una hipoteca de 135 000 dólares por una casa a 30 años y que los pagos máximos que puede realizar el cliente son de 1000 dólares mensuales. ¿Cuál será el interés más alto que podrá pagar?

24. El medicamento administrado a un paciente produce una concentración en la corriente sanguínea dada por $c(t) = Ate^{-t/3}$ miligramos por mililitro, t horas después de inyectarle A unidades. La máxima concentración segura es de 1 mg/ml.

   a. ¿Qué dosis deberá inyectársele al paciente para alcanzar la máxima concentración segura y cuándo se presenta esta concentración?

   b. Una cantidad adicional del medicamento deberá administrarse al paciente después de que la concentración disminuya a 0.25 mg/ml. Determine, con una aproximación al minuto más cercano, cuándo debe aplicarse la segunda inyección.

   c. Suponiendo que la concentración producida por inyecciones consecutivas es aditiva y que 75% de la dosis inyectada originalmente se administra en la segunda inyección, ¿cuándo será el momento de aplicar la tercera inyección?

25. Sea $f(x) = 3^{3x+1} - 7 \cdot 5^{2x}$.

   a. Use los comandos o instrucciones de `Maple solve` y `fsolve` para tratar de encontrar todos los ceros de f.

   b. Grafique $f(x)$ para obtener las aproximaciones iniciales de los ceros de f.

   c. Con el método de Newton encuentre los ceros de f con una exactitud de 10^{-16}.

   d. Encuentre algebraicamente las soluciones exactas de $f(x) = 0$.

26. Repita el ejercicio 25 usando $f(x) = 2^{x^2} - 3 \cdot 7^{x+1}$.

27. El modelo logístico del crecimiento demográfico se describe por medio de una ecuación de la forma

$$P(t) = \frac{P_L}{1 - ce^{-kt}},$$

donde P_L, c y $k > 0$ son constantes y $P(t)$ es la población en el tiempo t. P_L representa el valor límite de la población, ya que $\lim_{t\to\infty} P(t) = P_L$. Utilice los datos de los censos correspondientes a los años 1950, 1960 y 1970 que vienen en la tabla de la página 104 para determinar las constantes P_L, c y k para un modelo logístico de crecimiento. Utilice el modelo logístico para predecir la población de Estados Unidos en los años 1980 y 2010, suponiendo que $t = 0$ en 1950. Compare con el valor real la predicción relativa a 1980.

28. El modelo de Gompertz para el crecimiento demográfico se describe por medio de

$$P(t) = P_L e^{-ce^{-kt}},$$

donde P_L, c y $k > 0$ son constantes y $P(t)$ es la población en el tiempo t. Repita el ejercicio 27 usando el modelo de Gompertz en vez del modelo logístico.

29. El jugador A dejará en cero (por una puntuación de 21 a 0) al jugador B en un partido de raquetbol con una probabilidad de

$$P = \frac{1+p}{2}\left(\frac{p}{1-p+p^2}\right)^{21},$$

donde p denota la probabilidad de que A gane un intercambio de tiros (independientemente del servicio) (Véase [Keller, J, p. 267].) Determine, con una exactitud de 10^{-3}, el valor mínimo de p que garantice que A dejará en cero a B al menos en la mitad de los partidos que jueguen.

30. En el diseño de los vehículos para todo tipo de terreno, es necesario tener en cuenta las fallas cuando se trata de librar dos tipos de obstáculos. Una es la *falla por rozamiento*, y ocurre cuando el vehículo intenta cruzar un obstáculo que hace que su fondo toque el suelo. La otra recibe el nombre de *falla por colisión de la defensa delantera* y ocurre cuando el vehículo desciende por una zanja y la defensa delantera toca el suelo.

La figura anexa, adaptada de [Bek], muestra los componentes asociados al segundo tipo de falla. En ella se indica que el ángulo máximo α que puede alcanzar un vehículo cuando β es el ángulo máximo en que *no* ocurre la falla por rozamiento satisface la ecuación

$$A \operatorname{sen} \alpha \cos \alpha + B \operatorname{sen}^2 \alpha - C \cos \alpha - E \operatorname{sen} \alpha = 0,$$

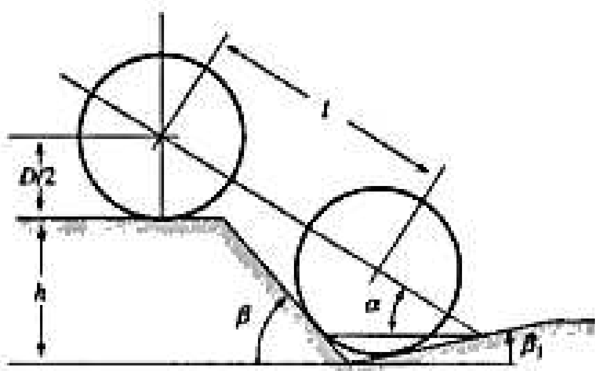
donde

$$A = l \operatorname{sen} \beta_1, \quad B = l \cos \beta_1, \quad C = (h + 0.5D)\operatorname{sen} \beta_1 - 0.5D \tan \beta_1,$$

y $\quad E = (h + 0.5D) \cos \beta_1 - 0.5D.$

a. Se afirma que, cuando $l = 89$ pulg, $h = 49$ pulg, $D = 55$ pulg y $\beta_1 = 11.5°$, el ángulo α será aproximadamente de 33°. Verifique este resultado.

b. Encuentre α para la situación en que l, h y β_1 son iguales como en la parte (a), pero $D = 30$ pulg.



2.4 Análisis de error para los métodos iterativos

En esta sección investigaremos el orden de convergencia de los esquemas de iteración funcional y, con el propósito de obtener una rápida convergencia, redescubriremos el método de Newton. También estudiaremos los métodos para acelerar la convergencia del método de Newton en circunstancias especiales. Ante todo, necesitamos un procedimiento nuevo para medir la rapidez con que converge una sucesión.

Definición 2.6 Supongamos que $\{p_n\}_{n=0}^{\infty}$ es una sucesión que converge a p, con $p_n \neq p$ para toda n. Si existen constantes positivas λ y α con

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lambda,$$

entonces $\{p_n\}_{n=0}^{\infty}$ **converge a p con orden α y una constante de error asintótica λ.** ■

Se dice que un método iterativo de la forma $p_n = g(p_{n-1})$ es de **orden** α, si la sucesión $\{p_n\}_{n=0}^{\infty}$ converge a la solución $p = g(p)$ con orden α.

En general, una sucesión con un alto orden de convergencia converge más rápidamente que una con un orden más bajo. La constante asintótica influye en la rapidez de convergencia, pero no es tan importante como el orden. Enfocaremos nuestra atención en dos casos de orden.

1. Si $\alpha = 1$, la sucesión será **linealmente convergente.**
2. Si $\alpha = 2$, la sucesión será **cuadráticamente convergente.**

En el siguiente ejemplo se compara una sucesión linealmente convergente con una cuadráticamente convergente, y se demuestra por qué trataremos de encontrar métodos que produzcan sucesiones convergentes de un orden superior.

EJEMPLO 1 Supongamos que $\{p_n\}_{n=0}^{\infty}$ converge linealmente a 0, con

$$\lim_{n\to\infty} \frac{|p_{n+1}|}{|p_n|} = 0.5$$

y que $\{\tilde{p}_n\}_{n=0}^{\infty}$ converge cuadráticamente a 0 con la misma constante de error asintótico,

$$\lim_{n\to\infty} \frac{|\tilde{p}_{n+1}|}{|\tilde{p}_n|^2} = 0.5.$$

Por razones de simplicidad, supongamos que

$$\frac{|p_{n+1}|}{|p_n|} \approx 0.5 \quad \text{y} \quad \frac{|\tilde{p}_{n+1}|}{|\tilde{p}_n|^2} \approx 0.5.$$

En el esquema linealmente convergente, esta suposición significa que

$$|p_n - 0| = |p_n| \approx 0.5|p_{n-1}| \approx (0.5)^2|p_{n-2}| \approx \cdots \approx (0.5)^n|p_0|,$$

mientras que el procedimiento cuadráticamente convergente tiene

$$\begin{aligned} |\tilde{p}_n - 0| = |\tilde{p}_n| &\approx 0.5|\tilde{p}_{n-1}|^2 \approx (0.5)[0.5|\tilde{p}_{n-2}|^2]^2 = (0.5)^3\,|\tilde{p}_{n-2}|^4 \\ &\approx (0.5)^3[(0.5)|\tilde{p}_{n-3}|^2]^4 = (0.5)^7|\tilde{p}_{n-3}|^8 \\ &\approx \cdots \approx (0.5)^{2^n-1}|\tilde{p}_0|^{2^n}. \end{aligned}$$

La tabla 2.7 ilustra la rapidez relativa de convergencia a cero de las sucesiones cuando $|p_0| = |\tilde{p}_0| = 1$.

Tabla 2.7

n	Sucesión lineal convergente $\{p_n\}_{n=0}^{\infty}$ $(0.5)^n$	Sucesión cuadrática convergente $\{\tilde{p}_n\}_{n=0}^{\infty}$ $(0.5)^{2^n-1}$
1	5.0000×10^{-1}	5.0000×10^{-1}
2	2.5000×10^{-1}	1.2500×10^{-1}
3	1.2500×10^{-1}	7.8125×10^{-3}
4	6.2500×10^{-2}	3.0518×10^{-5}
5	3.1250×10^{-2}	4.6566×10^{-10}
6	1.5625×10^{-2}	1.0842×10^{-19}
7	7.8125×10^{-3}	5.8775×10^{-39}

La sucesión cuadráticamente convergente se encuentra cerca de 0 a menos de 10^{-38} por el séptimo término. Se necesitan 126 términos por lo menos para garantizar esta precisión de la sucesión linealmente convergente.

En general, las sucesiones cuadráticamente convergentes lo hacen con mucha mayor rapidez que las que convergen sólo de modo lineal, pero muchas técnicas que generan sucesiones convergentes lo hacen sólo en forma lineal.

Teorema 2.7 Sea $g \in C[a, b]$ tal que $g(x) \in [a, b]$ para toda $x \in C[a, b]$. Supongamos, además, que g' es continua en (a, b) y que existe una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \qquad \text{para todo } x \in (a, b).$$

Si $g'(p) \neq 0$, entonces para cualquier número p_0 en $[a, b]$ la sucesión

$$p_n = g(p_{n-1}), \qquad \text{para} \quad n \geq 1,$$

converge sólo linealmente en el único punto fijo p en $[a, b]$. ■

Demostración En el teorema 2.3 del punto fijo que vimos en la sección 2.2 encontramos que la sucesión converge a p. Puesto que g' existe en $[a, b]$ podemos aplicar el teorema del valor medio a g para demostrar que, para cualquier n,

$$p_{n+1} - p = g(p_n) - g(p) = g'(\xi_n)(p_n - p),$$

donde ξ_n está entre p_n y p. Puesto que $\{p_n\}_{n=0}^{\infty}$ converge a p, $\{\xi_n\}_{n=0}^{\infty}$ también convergerá a p. Como g' es continua en $[a, b]$, tendremos

$$\lim_{n\to\infty} g'(\xi_n) = g'(p).$$

Por tanto,

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} = \lim_{n\to\infty} g'(\xi_n) = g'(p) \quad \text{y} \quad \lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = |g'(p)|.$$

En consecuencia, la iteración de punto fijo muestra una convergencia lineal con una constante de error asintótico $|g'(p)|$ siempre que $g'(p) \neq 0$. ■■■

El teorema 2.7 establece que, en el caso de los métodos de punto fijo, la convergencia de orden superior puede ocurrir sólo cuando $g'(p) = 0$. El resultado siguiente describe otras condiciones que garantizan la convergencia cuadrática que buscamos.

Teorema 2.8 Sea p una solución de la ecuación $x = g(x)$. Supongamos que $g'(p) = 0$ y g'' es continua y está estrictamente acotada por M en un intervalo abierto I que contiene a p. Entonces existe una $\delta > 0$ tal que, para $p_0 \in [p - \delta, p + \delta]$ la sucesión definida por $p_n = g(p_{n-1})$, cuando $n \geq 1$, converge al menos cuadráticamente a p. Además, para valores suficientemente grandes de n,

$$|p_{n+1} - p| < \frac{M}{2} |p_n - p|^2.$$ ■

Demostración Escoja k en $(0, 1)$ y $\delta > 0$ tal que en el intervalo $[p - \delta, p + \delta]$, contenido en I, tenemos $|g'(x)| \leq k$ y g'' sea continua. Dado que $|g'(x)| \leq k < 1$ el argumento empleado en la demostración del teorema 2.5 de la sección 2.3 indica que los términos de la sucesión $\{p_n\}_{n=0}^{\infty}$ están contenidos en $[p - \delta, p + \delta]$. Al desarrollar $g(x)$ en un polinomio lineal de Taylor para $x \in [p - \delta, p + \delta]$ obtenemos:

$$g(x) = g(p) + g'(p)(x - p) + \frac{g''(\xi)}{2}(x - p)^2,$$

donde ξ se encuentra entre x y p. Las hipótesis $g(p) = p$ y $g'(p) = 0$ significan que

$$g(x) = p + \frac{g''(\xi)}{2}(x - p)^2.$$

En particular, cuando $x = p_n$,

$$p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2}(p_n - p)^2,$$

con ξ_n entre p_n y p. Por tanto,

$$p_{n+1} - p = \frac{g''(\xi_n)}{2}(p_n - p)^2.$$

Puesto que $|g'(x)| \leq k < 1$ en $[p - \delta, p + \delta]$ y g manda $[p - \delta, p + \delta]$ en sí mismo, del teorema de punto fijo se deduce que $\{p_n\}_{n=0}^{\infty}$ converge a p. Pero ξ_n se encuentra entre p y p_n para cada n, de modo que $\{\xi_n\}_{n=0}^{\infty}$ también converge a p, y

$$\lim_{n \to \infty} \frac{|p_{n+1} - p|}{|p_n - p|^2} = \frac{|g''(p)|}{2}.$$

El resultado anterior implica que la sucesión $\{p_n\}_{n=0}^{\infty}$ es cuadráticamente convergente si $g''(p) \neq 0$ y con convergencia de orden superior si $g''(p) = 0$.

Puesto que g'' está estrictamente acotada por M en el intervalo $[p - \delta, p + \delta]$ esto también implica que, con valores suficientemente grandes de n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2.$$

■ ■ ■

Los teoremas 2.7 y 2.8 nos indican que nuestra búsqueda de los métodos de punto fijo cuadráticamente convergentes debería señalar hacia las funciones cuyas derivadas se anulan en el punto fijo.

La manera más fácil de plantear un problema de punto fijo relacionado con el de la búsqueda de raíces $f(x) = 0$ consiste en restar a x un múltiplo de $f(x)$. Por tanto, a continuación consideraremos un esquema de la forma

$$p_n = g(p_{n-1}), \qquad \text{para} \quad n \geq 1,$$

para g de la forma

$$g(x) = x - \phi(x) f(x),$$

donde ϕ es una función diferenciable que será elegida más tarde.

Para que el procedimiento iterativo derivado de g sea cuadráticamente convergente, es necesario tener $g'(p) = 0$ cuando $f(p) = 0$. Dado que

$$g'(x) = 1 - \phi'(x) f(x) - f'(x)\phi(x),$$

tendremos

$$g'(p) = 1 - \phi'(p)f(p) - f'(p)\,\phi(p) = 1 - \phi'(p)\cdot 0 - f'(p)\phi(p) = 1 - f'(p)\phi(p),$$

y $g'(p) = 0$ si y sólo si $\phi(p) = 1/f'(p)$.

Un enfoque razonable es suponer que $\phi(x) = 1/f'(x)$, lo cual garantizará que $\phi(p) = 1/f'(p)$ y entonces el procedimiento natural para producir la convergencia cuadrática será

$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}.$$

Éste, por supuesto, es el método de Newton.

En la explicación anterior, impusimos la restricción de que $f'(p) \neq 0$, donde p es la solución de $f(x) = 0$. Conforme a la definición del método de Newton, es evidente que pueden surgir dificultades si $f'(p_n)$ tiende a cero simultáneamente con $f(p_n)$. En particular, el método de Newton y el de la secante generalmente ocasionarán problemas si $f'(p) = 0$ cuando $f(p) = 0$. Para examinar más a fondo estas dificultades, damos la siguiente definición.

Definición 2.9 Una solución p de $f(x) = 0$ es un **cero de multiplicidad** m de f si para $x \neq p$, podemos escribir $f(x) = (x - p)^m\, q(x)$, donde $\lim_{x\to p} q(x) \neq 0$. ■

En esencia, $q(x)$ representa la parte de $f(x)$ que no contribuye al cero de f. El siguiente resultado proporciona un método fácil para identificar los ceros **simples** de una función, los que tienen multiplicidad uno.

Teorema 2.10 $f \in C^1[a, b]$ tiene un cero simple en p en (a, b) si y sólo si $f(p) = 0$, pero $f'(p) \neq 0$. ■

Demostración Si f tiene un cero simple en p, entonces $f(p) = 0$ y $f(x) = (x - p)q(x)$, donde $\lim_{x\to p} q(x) \neq 0$. Puesto que $f \in C^1[a, b]$,

$$f'(p) = \lim_{x\to p} f'(x) = \lim_{x\to p}\,[q(x) + (x - p)q'(x)] = \lim_{x\to p} q(x) \neq 0.$$

Por el contrario, si $f(p) = 0$, pero $f'(p) \neq 0$, desarrollamos f en el polinomio de Taylor de grado cero alrededor de p. Entonces

$$f(x) = f(p) + f'(\xi(x))(x - p) = (x - p)\,f'(\xi(x)),$$

donde $\xi(x)$ se encuentra entre x y p. Dado que $f \in C^1[a, b]$,

$$\lim_{x\to p} f'(\xi(x)) = f'\left(\lim_{x\to p} \xi(x)\right) = f'(p) \neq 0.$$

Definiendo $q = f' \circ \xi$ tenemos que $f(x) = (x - p)q(x)$, donde $\lim_{x\to p} q(x) \neq 0$. Por tanto, f tiene un cero simple en p. ■ ■ ■

En el ejercicio 10 se aplica la siguiente generalización del teorema 2.10.

Teorema 2.11 La función $f \in C^m[a, b]$ tiene un cero de multiplicidad m en p en (a, b) si y sólo si

$$0 = f(p) = f'(p) = f''(p) = \cdots = f^{(m-1)}(p), \quad \text{pero} \quad f^{(m)}(p) \neq 0. \quad ■$$

El resultado del teorema 2.10 implica que alrededor de p existe un intervalo tal que el método de Newton converge cuadráticamente en p para cualquier aproximación inicial $p_0 = p$, a condición de que p sea un cero simple. El siguiente ejemplo muestra cómo la convergencia cuadrática posiblemente no ocurra cuando el cero no es simple.

EJEMPLO 2 La función descrita por $f(x) = e^x - x - 1$ tiene un cero de multiplicidad dos en $p = 0$, porque $f(0) = e^0 - 0 - 1 = 0$ y $f'(0) = e^0 - 1 = 0$, pero $f''(0) = e^0 = 1$. De hecho, podemos expresar $f(x)$ en la forma

$$f(x) = (x - 0)^2 \frac{e^x - x - 1}{x^2},$$

donde, por la regla de L'Hôpital,

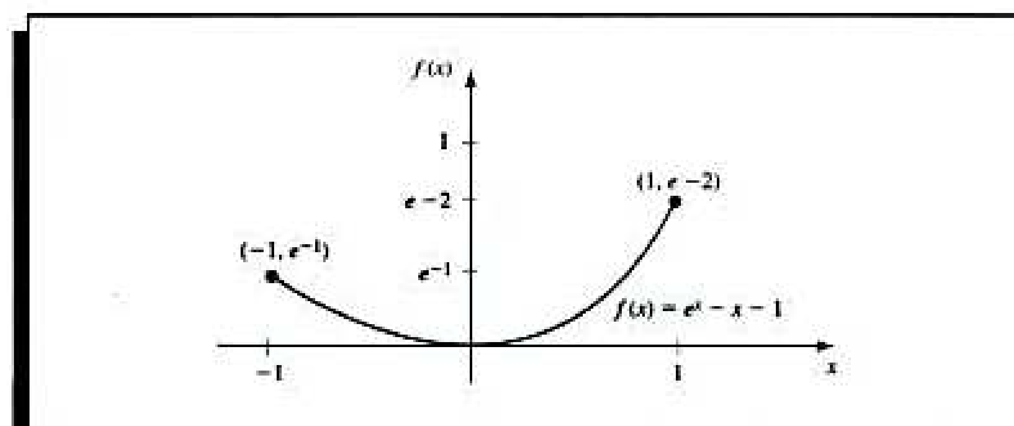
$$\lim_{x \to 0} \frac{e^x - x - 1}{x^2} = \lim_{x \to 0} \frac{e^x - 1}{2x} = \lim_{x \to 0} \frac{e^x}{2} = \frac{1}{2} \neq 0.$$

En la tabla 2.8 se incluyen los términos que se generaron con el método de Newton aplicado a f con $p_0 = 1$. La sucesión converge claramente a cero, pero no es cuadráticamente convergente. La gráfica de f se muestra en la figura 2.11. ■

Tabla 2.8

n	p_n	n	p_n
0	1.0	9	2.7750×10^{-3}
1	0.58198	10	1.3881×10^{-3}
2	0.31906	11	6.9411×10^{-4}
3	0.16800	12	3.4703×10^{-4}
4	0.08635	13	1.7416×10^{-4}
5	0.04380	14	8.8041×10^{-5}
6	0.02206	15	4.2610×10^{-5}
7	0.01107	16	1.9142×10^{-6}
8	0.005545		

Figura 2.11



Un método para resolver el problema de las raíces múltiples consiste en definir una función μ por medio de

$$\mu(x) = \frac{f(x)}{f'(x)}.$$

Si p es un cero de f de multiplicidad m y si $f(x) = (x - p)^m q(x)$, entonces

$$\mu(x) = \frac{(x-p)^m\, q(x)}{m(x-p)^{m-1}\, q(x) + (x-p)^m\, q'(x)}$$

$$= (x-p)\frac{q(x)}{mq(x) + (x-p)q'(x)}$$

también tiene un cero en p. Pero como $q(p) \neq 0$,

$$\frac{q\,(p)}{mq\,(p) + (p-p)q'(p)} = \frac{1}{m} \neq 0,$$

por tanto, p es un cero simple de μ. Así, podemos aplicar el método de Newton a la función μ para obtener

$$g(x) = x - \frac{\mu(x)}{\mu'(x)} = x - \frac{f(x)/f'(x)}{\{[f'(x)]^2 - [f(x)][f''(x)]\}/[f'(x)]^2}$$

o bien

$$g(x) = x - \frac{f(x)\,f'(x)}{[f'(x)]^2 - f(x)\,f''(x)}. \tag{2.11}$$

Si g tiene las condiciones de continuidad necesarias, la iteración funcional aplicada a g será cuadráticamente convergente sin importar la multiplicidad de la raíz de f. En teoría, el único inconveniente de este método es el cálculo adicional de $f''(x)$ y el procedimiento más laborioso con que se calculan las iteraciones. Sin embargo, en la práctica la presencia de un cero múltiple puede ocasionar serios problemas de redondeo porque el denominador de (2.11) consta de la diferencia de dos números que están cercanos a la raíz.

EJEMPLO 3 La tabla 2.9 contiene las aproximaciones de la raíz doble en $x = 0$ de $f(x) = e^x - x - 1$ utilizando $p_n = g(p_{n-1})$, para $n \geq 1$, donde g está dada por (2.11). Los resultados fueron obtenidos usando una calculadora con diez dígitos de precisión. Elegimos la aproximación inicial de $p_0 = 1$ de modo que las entradas puedan compararse con las de la tabla 2.8. Lo que no aparece en la tabla 2.9 es que no se logra mejoramiento alguno en la aproximación de la raíz $-2.8085217 \times 10^{-7}$ en los cálculos subsecuentes cuando se usa esta calculadora, ya que el denominador y el numerador se acercan al cero. ■

Tabla 2.9

n	p_n
1	$-2.3421061 \times 10^{-1}$
2	$-8.4582788 \times 10^{-3}$
3	$-1.1889524 \times 10^{-5}$
4	$-6.8638230 \times 10^{-6}$
5	$-2.8085217 \times 10^{-7}$

EJEMPLO 4 En el ejemplo 3 de la sección 2.2 encontramos la raíz $p = 1.36523001$ de $f(x) = x^3 + 4x^2 - 10 = 0$. Para comparar la convergencia para una raíz de multiplicidad uno por el método de Newton y el método modificado de Newton que se menciona en la ecuación (2.11), sea

$$\text{(i)} \qquad p_n = p_{n-1} - \frac{p_{n-1}^3 + 4p_{n-1}^2 - 10}{3p_{n-1}^2 + 8p_{n-1}}, \qquad \text{del método de Newton}$$

y, $p_n = g(p_{n-1})$, donde g se obtiene de la ecuación (2.11),

$$\text{(ii)} \qquad p_n = p_{n-1} - \frac{(p_{n-1}^3 + 4p_{n-1}^2 - 10)(3p_{n-1}^2 + 8p_{n-1})}{(3p_{n-1}^2 + 8p_{n-1})^2 - (p_{n-1}^3 + 4p_{n-1}^2 - 10)(6p_{n-1} + 8)}.$$

Cuando $p_0 = 1.5$, las tres primeras iteraciones para (i) y (ii) se incluyen en la tabla 2.10. Los resultados muestran la convergencia rápida de ambos métodos en el caso de una raíz simple. ■

Tabla 2.10

	(i)	(ii)
p_1	1.37333333	1.35689898
p_2	1.36526201	1.36519585
p_3	1.36523001	1.36523001

CONJUNTO DE EJERCICIOS 2.4

1. Use el método de Newton para encontrar las soluciones de los siguientes problemas con una exactitud de 10^{-5}.
   a. $x^2 - 2xe^{-x} + e^{-2x} = 0$, para $0 \le x \le 1$
   b. $\cos(x + \sqrt{2}) + x(x/2 + \sqrt{2}) = 0$, para $-2 \le x \le -1$
   c. $x^3 - 3x^2(2^{-x}) + 3x(4^{-x}) - 8^{-x} = 0$, para $0 \le x \le 1$
   d. $e^{6x} + 3(\ln 2)^2 e^{2x} - (\ln 8)\, e^{4x} - (\ln 2)^3 = 0$, para $-1 \le x \le 0$
2. Repita el ejercicio 1 aplicando el método modificado de Newton-Raphson descrito en la ecuación (2.11). ¿Mejora la rapidez o la exactitud en comparación con el ejercicio 1?
3. Aplique el método de Newton y el método modificado de Newton-Raphson descrito en la ecuación (2.11) para encontrar una solución del siguiente problema con una exactitud de 10^{-5}:

   $$e^{6x} + 1.441e^{2x} - 2.079e^{4x} - 0.3330 = 0 \quad \text{para} \quad -1 \le x \le 0.$$

   Este es el mismo problema que 1(d), sólo que el coeficiente ha sido reemplazado por sus aproximaciones de cuatro dígitos. Compare las soluciones con los resultados de 1(d) y de 2(d).
4. Demuestre que las sucesiones siguientes convergen linealmente a $p = 0$. ¿Qué tan grande debe ser n antes que $|p_n - p| \le 5 \times 10^{-2}$?
   a. $p_n = \dfrac{1}{n}, \quad n \ge 1$
   b. $p_n = \dfrac{1}{n^2}, \quad n \ge 1$
5. a. Demuestre que, para cualquier entero positivo k, la sucesión definida por $p_n = 1/n^k$ converge linealmente a $p = 0$.
   b. Para cada par de enteros k y m, determine un número N para el cual $1/N^k < 10^{-m}$.

6. a. Demuestre que la sucesión $p_n = 10^{-2^n}$ converge cuadráticamente en cero.
   b. Demuestre que la sucesión $p_n = 10^{-n^k}$ no converge cuadráticamente a cero, sin importar el tamaño del exponente $k > 1$.
7. a. Construya una sucesión que converja a cero de orden 3.
   b. Suponga que $\alpha > 1$. Construya una sucesión que converja a cero de orden α.
8. Suponga que p es una raíz de multiplicidad m de f donde f'' es continua en un intervalo abierto que contiene p. Demuestre que el siguiente método de punto fijo tiene $g'(p) = 0$:
$$g(x) = x - \frac{m f(x)}{f'(x)}.$$
9. Demuestre que el algoritmo de bisección 2.1 da una sucesión con una cota de error que converge linealmente a cero.
10. Suponga que f tiene m derivadas continuas. Modifique la demostración del teorema 2.10 para probar que f tiene una raíz de multiplicidad m en p si y sólo si
$$0 = f(p) = f'(p) = \cdots = f^{(m-1)}(p), \quad \text{pero} \quad f^{(m)}(p) \neq 0.$$
11. El método iterativo para resolver $f(x) = 0$, dado por el método de punto fijo $g(x) = x$, donde
$$p_n = g(p_{n-1}) = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})} - \frac{f''(p_{n-1})}{2f'(p_{n-1})}\left[\frac{f(p_{n-1})}{f'(p_{n-1})}\right]^2, \quad \text{para } n = 1, 2, 3, \ldots,$$
tiene $g'(p) = g''(p) = 0$. Esto generalmente producirá una convergencia cúbica ($\alpha = 3$). Use el análisis del ejemplo 1 para comparar la convergencia cuadrática y la convergencia cúbica.
12. Puede demostrarse (véase, por ejemplo, [DaB, pp. 228-229]) que, si $\{p_n\}_{n=0}^{\infty}$ son aproximaciones que convergen mediante el método de la secante a p, la solución de $f(x) = 0$, entonces existe una constante C con $|p_{n+1} - p| \approx C|p_n - p|\,|p_{n-1} - p|$ para valores suficientemente grandes de n. Suponga que $\{p_n\}$ converge a p con orden α, demuestre que $\alpha = (1 + \sqrt{5})/2$. (*Nota*: Ello significa que el orden de convergencia del método de la secante es aproximadamente 1.62).

2.5 Convergencia acelerada

Rara vez podemos darnos el lujo de tener una convergencia cuadrática; por ello, a continuación estudiaremos una técnica denominada **método Δ^2 de Aitken**, el cual sirve para acelerar la convergencia de una sucesión que sea linealmente convergente, prescindiendo de su origen o aplicación.

Supongamos que $\{p_n\}_{n=0}^{\infty}$ es una sucesión linealmente convergente con un límite p. Para impulsar la construcción de una sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ que converja más rápidamente a p que $\{p_n\}_{n=0}^{\infty}$, supongamos primero que los signos de $p_n - p$, $p_{n+1} - p$, y $p_{n+2} - p$ son iguales y que n es suficientemente grande como para que
$$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}.$$
Entonces,
$$(p_{n+1} - p)^2 \approx (p_{n+2} - p)(p_n - p),$$
por tanto,
$$p_{n+1}^2 - 2p_{n+1}p + p^2 \approx p_{n+2}p_n - (p_n + p_{n+2})p + p^2$$

y

$$(p_{n+2} + p_n - 2p_{n+1})\, p \approx p_{n+2}\, p_n - p_{n+1}^2.$$

Al despejar p obtenemos

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n}.$$

Si sumamos y restamos los términos p_n^2 y $2p_n p_{n+1}$ en el numerador, podemos reescribir esta expresión así:

$$\begin{aligned} p &\approx \frac{p_n^2 - p_n p_{n+2} - 2p_n p_{n+1} - 2\, p_n p_{n+1} - p_n^2 - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n} \\ &= \frac{(p_n^2 - p_n p_{n+2} + 2p_n p_{n+1}) - (p_n^2 - 2p_n\, p_{n+1} + p_{n+1}^2)}{p_{n+2} - 2p_{n+1} + p_n} \\ &= p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}. \end{aligned}$$

El método Δ^2 de Aitken se basa en la suposición de que la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ definida por

$$\hat{p}_n = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}, \qquad (2.12)$$

converge más rápidamente a p que la sucesión original $\{p_n\}_{n=0}^{\infty}$.

EJEMPLO 1 La sucesión $\{p_n\}_{n=1}^{\infty}$, donde $p_n = \cos(1/n)$, converge linealmente a $p = 1$. En la tabla 2.11 se incluyen los primeros términos de las sucesiones $\{p_n\}_{n=1}^{\infty}$ y $\{\hat{p}_n\}_{n=1}^{\infty}$. Sin duda parece que $\{\hat{p}_n\}_{n=1}^{\infty}$ converge más rápidamente a $p = 1$ que $\{p_n\}_{n=1}^{\infty}$. ■

Tabla 2.11

n	p_n	$\hat{p}_n$
1	0.54030	0.96178
2	0.87758	0.98213
3	0.94496	0.98979
4	0.96891	0.99342
5	0.98007	0.99541
6	0.98614	
7	0.98981	

La notación Δ asociada a esta técnica tiene su origen en la siguiente definición.

Definición 2.12 Dada la sucesión $\{p_n\}_{n=0}^{\infty}$, la **diferencia progresiva** Δp_n está definida por

$$\Delta p_n = p_{n+1} - p_n, \qquad \text{para } n \geq 0.$$

Las potencias mayores $\Delta^k p_n$ se definen recursivamente por medio de

$$\Delta^k p_n = \Delta(\Delta^{k-1} p_n), \quad \text{para } k \geq 2. \quad \blacksquare$$

La definición anterior significa que

$$\Delta^2 p_n = \Delta(p_{n+1} - p_n) = \Delta p_{n+1} - \Delta p_n = (p_{n+2} - p_{n+1}) - (p_{n+1} - p_n).$$

Por tanto,

$$\Delta^2 p_n = p_{n+2} - 2p_{n+1} + p_n,$$

y la fórmula $\hat{p}_n$ de la ecuación (2.12) puede escribirse así

$$\hat{p} = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}, \quad \text{para } n \geq 0. \tag{2.13}$$

Hasta ahora, al hablar del método Δ^2 de Aitken, hemos dicho que la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ converge a p más rápidamente que la sucesión original $\{p_n\}_{n=0}^{\infty}$, pero no hemos dicho lo que se entiende por una convergencia "más rápida". El teorema 2.13 explica y justifica esta terminología. La demostración del teorema se verá en el ejercicio 14.

Teorema 2.13 Supongamos que $\{p_n\}_{n=0}^{\infty}$ es una sucesión que converge linealmente al límite p y que, para todos los valores suficientemente grandes de n, tenemos $(p_n - p)(p_{n+1} - p) > 0$. Entonces, la sucesión $\{\hat{p}_n\}_{n=0}^{\infty}$ converge a p con mayor rapidez que $\{p_n\}_{n=0}^{\infty}$ en el sentido de que

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = 0. \quad \blacksquare$$

Al aplicar un método Δ^2 modificado de Aitken a una sucesión linealmente convergente obtenida mediante la iteración de punto fijo, podemos acelerar la convergencia a cuadrática. A este procedimiento se le conoce con el nombre de método de Steffensen, y difiere un poco de la aplicación del método Δ^2 de Aitken directamente a la sucesión de iteraciones de punto fijo que convergen linealmente. El método Δ^2 de Aitken deberá construir los términos en el orden:

$$p_0, \quad p_1 = g(p_0), \quad p_2 = g(p_1), \quad \hat{p}_0 = \{\Delta^2\}(p_0),$$
$$p_3 = g(p_2), \quad \hat{p}_1 = \{\Delta^2\}(p_1), \ldots,$$

donde $\{\Delta^2\}$ indica que se usa la ecuación (2.13). El método de Steffensen construye los mismos primeros cuatro términos p_0, p_1, p_2 y $\hat{p}_0$. No obstante, en este paso supone que $\hat{p}_0$ es una mejor aproximación de p que p_2 y aplica la iteración de punto fijo a $\hat{p}_0$ en vez de p_2. Al aplicar esta notación, la secuencia generada será

$$p_0^{(0)}, \quad p_1^{(0)} = g(p_0^{(0)}), \quad p_2^{(0)} = g(p_1^{(0)}), \quad p_0^{(1)} = \{\Delta^2\}(p_0^{(0)}), \quad p_1^{(1)} = g(p_0^{(1)}), \ldots$$

La ecuación (2.13) genera cada tercer término; los demás usan la iteración de punto fijo en el término anterior. El procedimiento se describe en el algoritmo 2.6

ALGORITMO 2.6

Método de Steffensen

Para encontrar una solución a $p = g(p)$ dada una aproximación inicial p_0:

ENTRADA aproximación inicial p_0; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Tome $i = 1$;

Paso 2 Mientras $i \leq N_0$ haga pasos 3-6.

Paso 3 Tome $p_1 = g(p_0)$; *(Calcule $p_1^{(i-1)}$.)*
$p_2 = g(p_1)$; *(Calcule $p_2^{(i-1)}$.)*
$p = p_0 - (p_1 - p_0)^2/(p_2 - 2p_1 + p_0)$. *(Calcule $p_0^{(i)}$.)*

Paso 4 Si $|p - p_0| < TOL$ entonces
SALIDA (p); *(Procedimiento terminado satisfactoriamente.)*
PARAR.

Paso 5 Tome $i = i + 1$.

Paso 6 Tome $p_0 = p$. (Redefina p_0.)

Paso 7 SALIDA ('El método falló después de N_0 iteraciones, N_0 = ', N_0);
(Procedimiento terminado sin éxito.)
PARAR. ■

Obsérvese que $\Delta^2 p_n$ puede ser cero, lo cual introduciría un cero en el denominador de la siguiente iteración. De ser así, terminaríamos la sucesión y escogeríamos $p_2^{(n-1)}$ como la respuesta aproximada.

EJEMPLO 2 Para resolver $x^3 + 4x^2 - 10 = 0$ mediante el método de Steffensen, sea $x^3 + 4x^2 = 10$ y despejamos x dividiendo entre $x + 4$. Con este procedimiento se produce el método de punto fijo

$$g(x) = \left(\frac{10}{x+4}\right)^{1/2},$$

utilizado en el ejemplo 3(d) de la sección 2.2.

El procedimiento de Steffensen con $p_0 = 1.5$ da los valores de la tabla 2.12. La exactitud de la iteración $p_0^{(2)} = 1.365230013$ es de nueve cifras decimales. En este ejemplo, con el método de Steffensen se obtuvo casi la misma rapidez de convergencia que con el método de Newton (véase el ejemplo 4 de la sección 2.4). ■

Tabla 2.12

k	$p_0^{(k)}$	$p_1^{(k)}$	$p_2^{(k)}$
0	1.5	1.348399725	1.367376372
1	1.365265224	1.365225534	1.365230583
2	1.365230013		

En el ejemplo 2, observamos que el método de Steffensen parece dar la convergencia cuadrática sin evaluar una derivada; el teorema 2.14 verifica que realmente es así. La demostración de este teorema se da en [He2, pp. 90-92] o en [IK, pp. 103-107].

Teorema 2.14 Supongamos que $x = g(x)$ tiene la solución p con $g'(p) \neq 1$. Si existe $\delta > 0$ tal que $g \in C^3[p - \delta, p + \delta]$, entonces con el método de Steffensen obtendremos la convergencia cuadrática para cualquier $p_0 \in [p - \delta, p + \delta]$. ■

CONJUNTO DE EJERCICIOS 2.5

1. Las siguientes sucesiones son linealmente convergentes. Genere los cinco primeros términos de la sucesión $\{\hat{p}_n\}$ por medio del método Δ^2 de Aitken.
   a. $p_0 = 0.5, \quad p_n = (2 - e^{p_{n-1}} + p_{n-1}^2)/3, \quad n \geq 1$
   b. $p_0 = 0.75, \quad p_n = (e^{p_{n-1}}/3)^{1/2}, \quad n \geq 1$
   c. $p_0 = 0.5, \quad p_n = 3^{-p_{n-1}}, \quad n \geq 1$
   d. $p_0 = 0.5, \quad p_n = \cos p_{n-1}, \quad n \geq 1$
2. Considere la función $f(x) = e^{6x} + 3(\ln 2)^2 e^{2x} - (\ln 8)e^{4x} - (\ln 2)^3$. Aplique el método de Newton con $p_0 = 0$ para aproximar una raíz de f. Genere términos hasta que $|p_{n+1} - p_n| < 0.0002$. Construya la sucesión $\{\hat{p}_n\}$. ¿Mejoró la convergencia?
3. Sean $g(x) = \cos(x - 1)$ y $p_0^{(0)} = 2$. Aplique el método de Steffensen para encontrar $p_0^{(1)}$.
4. Sean $g(x) = 1 + (\operatorname{sen} x)^2$ y $p_0^{(0)} = 1$. Aplique el método de Steffensen para encontrar $p_0^{(1)}$ y $p_0^{(2)}$.
5. El método de Steffensen se aplica a una función g por medio de $p_0^{(0)} = 1$ y $p_2^{(0)} = 3$ para obtener $p_0^{(1)} = 0.75$. ¿Qué podría ser $p_1^{(0)}$?
6. El método de Steffensen se aplica a una función g por medio de $p_0^{(0)} = 1$ y $p_1^{(0)} = \sqrt{2}$ para obtener $p_0^{(1)} = 2.7802$. ¿Qué es $p_2^{(0)}$?
7. Resuelva $x^3 - x - 1 = 0$ para la raíz en $[1, 2]$ con una exactitud de 10^{-4} aplicando el método de Steffensen y compare después los resultados con los del ejercicio 6 de la sección 2.2.
8. Resuelva $x - 2^{-x} = 0$ para la raíz en $[0, 1]$ con una exactitud de 10^{-4} empleando el método de Steffensen y compare después los resultados con los del ejercicio 8 de la sección 2.2.
9. Aplique el método de Steffensen con $p_0 = 2$ para calcular la aproximación de $\sqrt{3}$ con una exactitud de 10^{-4}. Después compare este resultado con los obtenidos en el ejercicio 9 de la sección 2.2 y en el ejercicio 10 de la sección 2.1.
10. Aplique el método de Steffensen para aproximar las soluciones de las siguientes ecuaciones con una exactitud de 10^{-5}.
    a. $x = (2 - e^x + x^2)/3$, donde g es la función en el ejercicio 11(a) de la sección 2.2.
    b. $x = 0.5(\operatorname{sen} x + \cos x)$, donde g es la función en el ejercicio 11(f) de la sección 2.2.
    c. $3x^2 - e^x = 0$, donde g es la función en el ejercicio 12(a) de la sección 2.2.
    d. $x - \cos x = 0$, donde g es la función en el ejercicio 12(b) de la sección 2.2.
11. Las sucesiones siguientes convergen a 0. Use el método Δ^2 de Aitken para generar $\{\hat{p}_n\}$ hasta que $|\hat{p}_n| \leq 5 \times 10^{-2}$:
    a. $p_n = \dfrac{1}{n}, \quad n \geq 1$
    b. $p_n = \dfrac{1}{n^2}, \quad n \geq 1$

12. Se dice que una sucesión $\{p_n\}$ es **superlinealmente convergente** a p si

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = 0.$$

a. Demuestre que, si $p_n \to p$ con orden α para $\alpha > 1$, entonces $\{p_n\}$ será superlinealmente convergente a p.

b. Demuestre que $p_n = \frac{1}{n^n}$ es superlinealmente convergente a 0, pero que no converge a cero con orden α para toda $\alpha > 1$.

13. Suponga que $\{p_n\}$ es superlinealmente convergente en p. Demuestre que

$$\lim_{n\to\infty} \frac{|p_{n+1} - p_n|}{|p_n - p|} = 1.$$

14. Demuestre el teorema 2.13 [*Sugerencia*: suponga que $\delta_n = (p_{n+1} - p)/(p_n - p) - \lambda$ y demuestre que $\lim_{n\to\infty} \delta_n = 0$. Después exprese $(\hat{p}_{n+1} - p)/(p_n - p)$ en función de δ_n, δ_{n+1} y λ.]

15. Sea $P_n(x)$ el polinomio de Taylor de grado n para $f(x) = e^x$ desarrollado alrededor de $x_0 = 0$.

a. Con x fija, demuestre que $p_n = P_n(x)$ satisface las hipótesis del teorema 2.13.

b. Sea $x = 1$. Use el método Δ^2 de Aitken para generar la sucesión $\hat{p}_0, \ldots, \hat{p}_8$.

c. En esta situación, ¿se acelera la convergencia con el método de Aitken?

2.6 Ceros de polinomios y el método de Müller

Un *polinomio de grado n* tiene la forma

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

donde las a_i, denominadas *coeficientes* de P, son constantes y $a_n \neq 0$. La función cero, $P(x) = 0$ para todos los valores de x, se considera un polinomio, pero sin que se le asigne grado alguno.

Teorema 2.15 **(Teorema fundamental de álgebra)**
Si $P(x)$ es un polinomio de grado $n \geq 1$ con coeficientes reales o complejos, entonces $P(x) = 0$ tiene al menos una raíz (posiblemente compleja). ■

Aunque el teorema 2.15 es básico en el estudio de las funciones elementales, la demostración habitual requiere técnicas tomadas del estudio de la teoría de las funciones complejas. Le recomendamos al lector consultar [SaS, p. 155], a fin de concluir una exposición sistemática de los temas necesarios para demostrar el teorema 2.15.

Una consecuencia importante de ese teorema es el corolario siguiente.

Corolario 2.16 Si $P(x)$ es un polinomio de grado $n \geq 1$ con coeficientes reales o complejos, entonces existen constantes únicas $x_1, x_2, \ldots, x_k$, posiblemente complejas, y enteros positivos $m_1, m_2, \ldots, m_k$ tales que $\sum_{i=1}^{k} m_i = n$ y

$$P(x) = a_n(x - x_1)^{m_1}(x - x_2)^{m_2} \cdots (x - x_k)^{m_k}.$$ ■

El corolario 2.16 establece que el conjunto de ceros de un polinomio es único y que, si cada cero x_i se cuenta el mismo número de veces que su multiplicidad m_i, entonces un polinomio de grado n tendrá exactamente n ceros.

El siguiente corolario del teorema fundamental de álgebra se usará con frecuencia en esta sección y en capítulos posteriores.

Corolario 2.17 Sean $P(x)$ y $Q(x)$ polinomios de grado a lo más n. Si $x_1, x_2, \ldots, x_k$, con $k > n$, son números distintos con $P(x_i) = Q(x_i)$ para $i = 1, 2, \ldots, k$, entonces $P(x) = Q(x)$, para todos los valores de x. ■

Si queremos localizar los ceros aproximados de un polinomio $P(x)$ con el procedimiento de Newton, necesitamos evaluar $P(x)$ en valores especificados. Puesto que $P(x)$ y $P'(x)$ son polinomios, la eficiencia computacional requiere evaluar estas funciones en la forma anidada que explicamos en la sección 1.2. El método de Horner incorpora esta técnica y, por lo mismo, requiere sólo n multiplicaciones y n sumas para evaluar un polinomio arbitrario de grado n.

Teorema 2.18 **(Método de Horner)**

Sea

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Si $b_n = a_n$ y

$$b_k = a_k + b_{k+1}x_0, \quad \text{para } k = n-1, n-2, \ldots, 1, 0,$$

entonces $b_0 = P(x_0)$. Más aún, si

$$Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2 x + b_1,$$

entonces

$$P(x) = (x - x_0)Q(x) + b_0. \qquad ■$$

Demostración Según la definición de $Q(x)$,

$$\begin{aligned}(x - x_0)Q(x) + b_0 &= (x - x_0)(b_n x^{n-1} + \cdots + b_2 x + b_1) + b_0\\ &= (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_2 x^2 + b_1 x)\\ &\quad - (b_n x_0 x^{n-1} + \cdots + b_2 x_0 x + b_1 x_0) + b_0\\ &= b_n x^n + (b_{n-1} - b_n x_0)x^{n-1} + \cdots + (b_1 - b_2 x_0)x + (b_0 - b_1 x_0).\end{aligned}$$

De acuerdo con la hipótesis, $b_n = a_n$ y $b_k - b_{k+1}x_0 = a_k$, por tanto

$$(x - x_0)Q(x) + b_0 = P(x) \quad \text{y} \quad b_0 = P(x_0). \qquad ■ ■ ■$$

EJEMPLO 1 Aplique el método de Horner para evaluar $P(x) = 2x^4 - 3x^2 + 3x - 4$ en $x_0 = -2$.

Cuando realizamos manualmente los cálculos en el método de Horner, primero construimos una tabla que sugiere el nombre de *división sintética* comúnmente aplicado a esta técnica. En este problema, la tabla es la siguiente:

	Coeficiente de x^4	Coeficiente de x^3	Coeficiente de x^2	Coeficiente de x	Término constante
$x_0 = -2$	$a_4 = 2$	$a_3 = 0$	$a_2 = -3$	$a_1 = 3$	$a_0 = -4$
		$b_4x_0 = -4$	$b_3x_0 = 8$	$b_2x_0 = -10$	$b_1x_0 = 14$
	$b_4 = 2$	$b_3 = -4$	$b_2 = 5$	$b_1 = -7$	$b_0 = 10$

Por tanto,

$$P(x) = (x + 2)(2x^3 - 4x^2 + 5x - 7) + 10. \qquad \blacksquare$$

Una ventaja más del uso del procedimiento de Horner (o división sintética) consiste en que, como

$$P(x) = (x - x_0)Q(x) + b_0,$$

donde

$$Q(x) = b_nx^{n-1} + b_{n-1}x^{n-2} + \cdots + b_2x + b_1,$$

al derivar respecto a x obtenemos

$$P'(x) = Q(x) + (x - x_0)Q'(x) \quad \text{y} \quad P'(x_0) = Q(x_0). \tag{2.14}$$

Cuando usamos el método de Newton–Raphson para encontrar un cero aproximado de un polinomio, podemos evaluar de la misma manera $P(x)$ y $P'(x)$.

EJEMPLO 2 Encuentre una aproximación a uno de los ceros de

$$P(x) = 2x^4 - 3x^2 + 3x - 4,$$

usando el procedimiento de Newton y la división sintética para evaluar $P(x_n)$ y $P'(x_n)$ en cada iteración x_n.

Con $x_0 = -2$ como aproximación inicial, obtuvimos $P(-2)$ en el ejemplo 1 por medio de

$x_0 = -2$	2	0	−3	3	−4	
	−4	8	−10	14		
	2	−4	5	−7	10	$= P(-2)$.

Usando el teorema 2.18 y la ecuación (2.14),

$$Q(x) = 2x^3 - 4x^2 + 5x - 7 \quad \text{y} \quad P'(-2) = Q(-2),$$

de modo que $P'(-2)$ puede encontrarse al evaluar $Q(-2)$ de manera similar:

$x_0 = -2$	2	−4	5	−7	
		−4	16	−42	
	2	−8	21	−49	$= Q(-2) = P'(-2)$

y

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = -2 - \frac{10}{-49} \approx -1.796.$$

Al repetir el procedimiento para encontrar x_2,

-1.796	2	0	-3	3	-4	
		-3.592	6.451	-6.197	5.742	
	2	-3.592	3.451	-3.197	1.742	$= P(x_1)$
		-3.592	12.902	-29.368		
	2	-7.184	16.353	-32.565	$= Q(x_1)$	$= P'(x_1)$.

Por tanto, $P(-1.796) = 1.742$, $P'(-1.796) = -32.565$, y

$$x_2 = -1.796 - \frac{1.742}{-32.565} \approx -1.7425.$$

De modo semejante, $x_3 = -1.73897$. Un cero real con cinco cifras decimales es -1.73896.

■

Obsérvese que el polinomio $Q(x)$ depende de la aproximación que se emplea, y cambia de una iteración a otra.

El algoritmo 2.7 calcula $P(x_0)$ y $P'(x_0)$ por medio del método de Horner.

ALGORITMO 2.7

Método de Horner

Para evaluar el polinomio

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = (x - x_0)Q(x) + b_0$$

y su derivada en x_0:

ENTRADA grado n; coeficientes $a_0, a_1, \ldots, a_n$; x_0.

SALIDA $y = P(x_0)$; $z = P'(x_0)$.

Paso 1 Tome $y = a_n$; (*Calcule b_n para P.*)
$z = a_n$. (*Calcule b_{n-1} para Q.*)

Paso 2 Para $j = n-1, n-2, \ldots, 1$
tome $y = x_0 y + a_j$; (*Calcule b_j para P.*)
$z = x_0 z + y$. (*Calcule b_{j-1} para Q.*)

Paso 3 Tome $y = x_0 y + a_0$. (*Calcule b_0 para P.*)

Paso 4 SALIDA (y, z);
PARAR.

■

Si la iteración N, x_N, en el procedimiento de Newton es un cero aproximado de P, entonces

$$P(x) = (x - x_N)Q(x) + b_0 = (x - x_N)Q(x) + P(x_N) \approx (x - x_N)Q(x).$$

de modo que $x - x_N$ será un factor aproximado de $P(x)$. Suponiendo que $\hat{x}_1 = x_N$ sea el cero aproximado de P y que $Q_1(x) \equiv Q(x)$ sea el factor aproximado, obtendremos

$$P(x) \approx (x - \hat{x}_1)Q_1(x).$$

Si aplicamos el método de Newton a $Q_1(x)$ podemos encontrar un segundo cero aproximado de P. Si $P(x)$ es un polinomio de grado n con n ceros reales, aplicamos varias veces este procedimiento para finalmente obtener $(n - 2)$ ceros aproximados de P y un factor cuadrático aproximado $Q_{n-2}(x)$. En esta etapa, podemos resolver $Q_{n-2}(x) = 0$ mediante una fórmula cuadrática para obtener los dos últimos ceros aproximados de P. Aunque podemos usar este método para obtener todos los ceros aproximados, se basa en el uso repetido de aproximaciones y puede generar resultados muy imprecisos.

El procedimiento que acabamos de describir se llama **deflación**. El problema de la exactitud de la deflación se debe al hecho de que, cuando obtenemos los ceros aproximados de $P(x)$, el método de Newton se aplica al polinomio reducido $Q_k(x)$, es decir, al que tiene la propiedad de que

$$P(x) \approx (x - \hat{x}_1)(x - \hat{x}_2) \cdots (x - \hat{x}_k)\, Q_k(x).$$

Un cero aproximado $\hat{x}_{k+1}$ de Q_k generalmente no aproximará una raíz de $P(x) = 0$ tan bien como una raíz de la ecuación reducida $Q_k(x) = 0$, y la inexactitud se incrementa al aumentar k. Una forma de superar esta dificultad consiste en utilizar las ecuaciones reducidas para obtener las aproximaciones $\hat{x}_2, \hat{x}_3, \ldots, \hat{x}_k$ a los ceros de P y mejorarlas después aplicando el método de Newton al polinomio original $P(x)$.

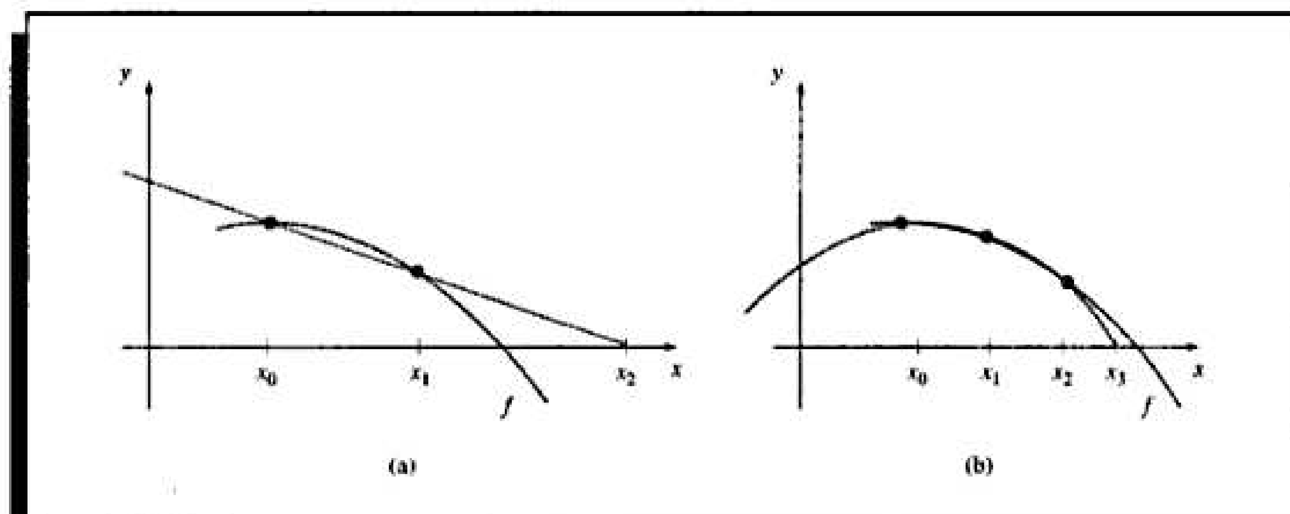
Un problema que se presenta al aplicar el método de Newton a los polinomios, es la posibilidad de que el polinomio contenga raíces complejas, cuando todos los coeficientes son números reales. Si la aproximación inicial mediante el método de Newton es un número real, también lo serán las aproximaciones subsecuentes. Una manera de superar esta dificultad consiste en comenzar con una aproximación inicial compleja y efectuar todos los cálculos por medio de la aritmética compleja. Otra manera se basa en el siguiente teorema.

Teorema 2.19 Si $z = a + bi$ es un cero complejo de multiplicidad m del polinomio $P(x)$, entonces $\bar{z} = a - bi$ también será un cero de multiplicidad m del polinomio $P(x)$ y $(x^2 - 2ax + a^2 + b^2)^m$ será un factor de $P(x)$. ∎

Podemos idear una división sintética que contenga polinomios cuadráticos para factorizar aproximadamente el polinomio, de modo que un término sea un polinomio cuadrático cuyas raíces complejas sean aproximaciones a las raíces del polinomio original. Esta técnica se describió con cierto detalle en la segunda edición del libro [BFR]. En vez de proceder en esa forma, ahora estudiaremos un método que fue propuesto inicialmente por D.E. Müller [Mu]. Podemos aplicarlo en cualquier problema de búsqueda de raíces, pero resulta de gran utilidad sobre todo al aproximar las raíces de polinomios.

El método de Müller es una extensión del método de la secante. Este último comienza con dos aproximaciones iniciales x_0 y x_1 y determina la siguiente aproximación x_2 como la intersección del eje x con la línea que cruza $(x_0, f(x_0))$ y $(x_1, f(x_1))$. (Véase la Fig. 2.12(a).) El método de Müller utiliza tres aproximaciones iniciales, x_0, x_1, y x_2 y determina la siguiente aproximación x_3 al considerar la intersección del eje x con la parábola que atraviese $(x_0, f(x_0))$, $(x_1 f(x_1))$ y $(x_2, f(x_2))$. (Véase la Fig. 2.12(b).)

Figura 2.12



La deducción del método de Müller comienza considerando el polinomio cuadrático

$$P(x) = a(x - x_2)^2 + b(x - x_2) + c$$

que pasa por $(x_0, f(x_0))$, $(x_1, f(x_1))$ y $(x_2, f(x_2))$. Podemos determinar las constantes, a, b y c a partir de las condiciones

$$f(x_0) = a(x_0 - x_2)^2 + b(x_0 - x_2) + c, \tag{2.15}$$

$$f(x_1) = a(x_1 - x_2) + b(x_1 - x_2) + c, \tag{2.16}$$

y

$$f(x_2) = a \cdot 0^2 + b \cdot 0 + c = c \tag{2.17}$$

para ser

$$c = f(x_2), \tag{2.18}$$

$$b = \frac{(x_0 - x_2)^2[f(x_1) - f(x_2)] - (x_1 - x_2)^2\,[f(x_0) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)}, \tag{2.19}$$

y

$$a = \frac{(x_1 - x_2)[f(x_0) - f(x_2)] - (x_0 - x_2)[f(x_1) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)}.$$

Si queremos determinar x_3, un cero de P, aplicamos la fórmula cuadrática a $P(x) = 0$. Sin embargo, debido a los problemas del error de redondeo ocasionados por la sustracción de números casi iguales, utilizaremos la fórmula como se indica en el ejemplo 5 de la sección 1.2:

$$x_3 - x_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

Esta fórmula ofrece dos posibilidades de x_3, según el signo que precede al término radical. En el método de Müller, el signo se elige de modo que corresponda al signo de b. De esa forma el denominador será el de mayor magnitud y hará que x_3 sea seleccionada como la raíz de P que está más cercana a x_2. Por tanto,

$$x_3 = x_2 - \frac{2c}{b + \text{signo}(b)\sqrt{b^2 - 4ac}},$$

donde a, b y c están dadas en la ecuación (2.15).

Una vez que determinamos x_3, reinicializamos el procedimiento usando x_1, x_2 y x_3 en vez de x_0, x_1 y x_2 para obtener la siguiente aproximación, x_4. El método prosigue hasta que se logra una conclusión satisfactoria. En cada paso el método contiene el radical $\sqrt{b^2 - 4ac}$, por tanto, puede aproximar las raíces complejas cuando $b^2 - 4ac < 0$. Con el algoritmo 2.8 se establece este procedimiento.

ALGORITMO 2.8

Método de Müller

Para obtener una solución de $f(x) = 0$ dadas tres aproximaciones, x_0, x_1 y x_2:

ENTRADA x_0, x_1, x_2; tolerancia TOL; número máximo de iteraciones N_0.

SALIDA solución aproximada p o mensaje de falla.

Paso 1 Tome $h_1 = x_1 - x_0$;
$h_2 = x_2 - x_1$;
$\delta_1 = (f(x_1) - f(x_0))/h_1$;
$\delta_2 = (f(x_2) - f(x_1))/h_2$;
$d = (\delta_2 - \delta_1)/(h_2 + h_1)$;
$i = 3$.

Paso 2 Mientras $i \leq N_0$ haga pasos 3-7.

Paso 3 $b = \delta_2 + h_2 d$;
$D = (b^2 - 4f(x_2)d)^{1/2}$. (*Nota: se puede necesitar aritmética compleja.*)

Paso 4 Si $|b - D| < |b + D|$ entonces tome $E = b + D$
si no, tome $E = b - D$.

Paso 5 Tome $h = -2f(x_2)/E$;
$p = x_2 + h$.

Paso 6 Si $|h| < TOL$ entonces
SALIDA (p); (*Procedimiento terminado satisfactoriamente.*)
PARAR.

Paso 7 Tome $x_0 = x_1$; (*Prepárese para la siguiente iteración.*)
$x_1 = x_2$;
$x_2 = p$;
$h_1 = x_1 - x_0$;
$h_2 = x_2 - x_1$;
$\delta_1 = (f(x_1) - f(x_0))/h_1$;

$\delta_2 = (f(x_2) - f(x_1))/h_2;$
$d = (\delta_2 - \delta_1)/(h_2 + h_1);$
$i = i + 1.$

Paso 8 SALIDA ('El método falló después de N_0 iteraciones, N_0 = ', N_0);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

EJEMPLO 3 Considere el polinomio $f(x) = 16x^4 - 40x^3 + 5x^2 + 20x + 6$. Al utilizar el algoritmo 2.8 con $TOL = 10^{-5}$ y diversos valores de x_0, x_1 y x_2, se obtienen los resultados que se proporcionan en la tabla 2.13.

Tabla 2.13

a.

i	$x_0 = 0.5$, $x_1 = -0.5$, $x_2 = 0$ x_i	$f(x_i)$
3	$-0.555556 + 0.598352i$	$-29.4007 - 3.89872i$
4	$-0.435450 + 0.102101i$	$1.33223 - 1.19309i$
5	$-0.390631 + 0.141852i$	$0.375057 - 0.670164i$
6	$-0.357699 + 0.169926i$	$-0.146746 - 0.00744629i$
7	$-0.356051 + 0.162856i$	$-0.183868 \times 10^{-2} + 0.539780 \times 10^{-3}i$
8	$-0.356062 + 0.162758i$	$0.286102 \times 10^{-5} + 0.953674 \times 10^{-6}i$

b.

i	$x_0 = 0.5$, $x_1 = 1.0$, $x_2 = 1.5$ x_i	$f(x_i)$
3	1.28785	-1.37624
4	1.23746	0.126941
5	1.24160	0.219440×10^{-2}
6	1.24168	0.257492×10^{-4}
7	1.24168	0.257492×10^{-4}

c.

i	$x_0 = 2.5$, $x_1 = 2.0$, $x_2 = 2.25$ x_i	$f(x_i)$
3	1.96059	-0.611255
4	1.97056	0.748825×10^{-2}
5	1.97044	-0.295639×10^{-4}
6	1.97044	-0.259639×10^{-4}

Usamos Maple para generar la parte (c) de la tabla 2.13. Para esto, definimos $f(x)$ y las aproximaciones iniciales como

```
>f:=x->16*x^4-40*x^3+5*x^2+20*x+6;
>p0:=0.5; p1:=-0.5; p2:=0.0;
```

Evaluamos el polinomio en los valores iniciales

```
>f0:=f(p0); f1:=f(p1); f2:=f(p2);
```

y obtenemos $c = 6$, $b = 10$, $a = 9$ y $p_3 = -0.5555555558 + 0.5983516452i$ usando las fórmulas del método de Müller:

```
>c:=f2;
>b:=((p0-p2)^2*(f1-f2)-(p1-p2)^)2*(f0-f2))/((p0-p2)*(p1-p2)*(p0-p1));
>a:=((p1-p2)*(f0-f2)-(p0-p2)*(f1-f2))/((p0-p2)*(p1-p2)*(p0-p1));
>p3:=p2-(2*c)/(b+(b/abs(b))*sqrt(b^2-4*a*c));
```

Generamos el valor p_3 usando aritmética compleja, al igual que en el cálculo

```
>f3:=f(p3);
```

lo que da $f_3 = -29.40070112 - 3.898724738i$.

Los valores reales de las raíces de la ecuación son 1.241677, 1.970446 y $-0.356062 \pm 0.162758i$, lo cual demuestra la exactitud de las aproximaciones obtenidas con el método de Müller. ■

En el ejemplo 3 se muestra que el método de Müller permite aproximar las raíces de los polinomios con varios valores iniciales. De hecho, este método generalmente converge a la raíz de un polinomio con cualquier aproximación inicial, aunque podemos construir problemas en que no haya convergencia en algunas elecciones de las aproximaciones iniciales. Esto puede suceder, por ejemplo, si para alguna i tenemos $f(x_i) = f(x_{i+1}) = f(x_{i+2}) \neq 0$. Entonces la ecuación cuadrática se reduce a una función constante no cero y nunca cruzará el eje x. Sin embargo, rara vez es así y los paquetes de computación que utiliza el método de Müller piden sólo una aproximación inicial por raíz e incluso la proporcionan como opción.

CONJUNTO DE EJERCICIOS 2.6

1. Obtenga las aproximaciones, con una exactitud de 10^{-4} a todos los ceros reales del siguiente polinomio aplicando el método de Newton.
   a. $f(x) = x^3 - 2x^2 - 5$
   b. $f(x) = x^3 + 3x^2 - 1$
   c. $f(x) = x^3 - x - 1$
   d. $f(x) = x^4 + 2x^2 - x - 3$
   e. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$
   f. $f(x) = x^5 - x^4 + 2x^3 - 3x^2 + x - 4$
2. Obtenga aproximaciones con un grado de exactitud de 10^{-5} a todos los ceros de los siguientes polinomios, encontrando primero los ceros reales mediante el método de Newton y reduciendo luego los polinomios de menor grado para determinar los ceros complejos.

a. $f(x) = x^4 + 5x^3 - 9x^2 - 85x - 136$

b. $f(x) = x^4 - 2x^3 - 12x^2 + 16x - 40$

c. $f(x) = x^4 + x^3 + 3x^2 + 2x + 2$

d. $f(x) = x^5 + 11x^4 - 21x^3 - 10x^2 - 21x - 5$

e. $f(x) = 16x^4 + 88x^3 + 159x^2 + 76x - 240$

f. $f(x) = x^4 - 4x^2 - 3x + 5$

g. $f(x) = x^4 - 2x^3 - 4x^2 + 4x + 4$

h. $f(x) = x^3 - 7x^2 + 14x - 6$

3. Repita el ejercicio 1 aplicando el método de Müller.

4. Repita el ejercicio 2 aplicando el método de Müller.

5. Use el método de Newton para obtener, con una exactitud de 10^{-3} los ceros y los puntos críticos de las siguientes funciones. Use esta información para trazar la gráfica de f.

a. $f(x) = x^3 - 9x^2 + 12$ b. $f(x) = x^4 - 2x^3 - 5x^2 + 12x - 5$

6. $f(x) = 10x^3 - 8.3x^2 + 2.295x - 0.21141 = 0$ tiene una raíz en $x = 0.29$. Use el método de Newton con una aproximación inicial $x_0 = 0.28$ para tratar de obtener esta raíz. Explique lo que sucede.

7. Use Maple para encontrar las raíces exactas del polinomio $f(x) = x^3 + 4x - 4$.

8. Use Maple para encontrar las raíces exactas del polinomio $f(x) = x^3 - 2x - 5$.

9. Aplique los métodos siguientes para obtener una solución con una exactitud de 10^{-4} para el problema

$$600x^4 - 550x^3 + 200x^2 - 20x - 1 = 0.$$

a. Método de bisección

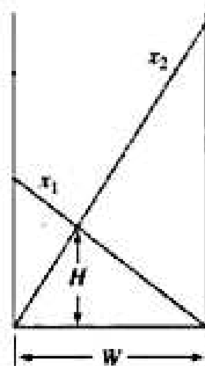
b. Método de Newton

c. Método de la secante

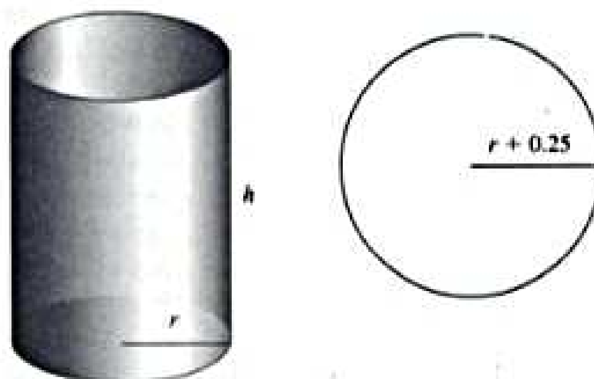
d. Método de la posición falsa

e. Método de Müller

10. Dos escaleras se cruzan en un pasillo de ancho W. Cada una llega de la base de un muro a un punto en el muro de enfrente. Las escaleras se cruzan a una altura H arriba del pavimiento. Dado que las longitudes de las escaleras son $x_1 = 20$ pies y $x_2 = 30$ pies y que $H = 8$ pies, calcule W.



11. Debemos fabricar una lata de forma cilíndrica circular recta que contenga 1000 cm³. La tapa circular de la parte superior y del fondo deben tener un radio de 0.25 cm más que el radio de la lata, para que el sobrante se utilice para sellar con la parte lateral. La hoja de material con que se construye esta parte de la lata también debe ser 0.25 cm más grande que la circunferencia de la lata, de modo que pueda hacerse un sello. Calcule, con una exactitud de 10^{-4}, la cantidad mínima de material necesaria para fabricar la lata.



12. En 1224, Leonardo de Pisa, mejor conocido como Fibonacci, resolvió el reto matemático de Juan de Palermo en presencia del emperador Federico II. El reto consistía en obtener una raíz de la ecuación $x^3 + 2x^2 + 10x = 20$. Primero demostró que la ecuación carecía de raíces racionales y de una raíz irracional euclidiana, es decir, no tenía ninguna raíz de una de las formas $a \pm \sqrt{b}$, $\sqrt{a} \pm \sqrt{b}$, $\sqrt{a \pm \sqrt{b}}$, o $\sqrt{\sqrt{a} \pm \sqrt{b}}$, donde a y b son números racionales. Después aproximó la única raíz real, probablemente aplicando un método algebraico de Omar Khayyam que incluía la intersección de un círculo y de una parábola. Su respuesta la dio en un sistema numérico de base 60 así:

$$1 + 22\left(\frac{1}{60}\right) + 7\left(\frac{1}{60}\right)^2 + 42\left(\frac{1}{60}\right)^3 + 33\left(\frac{1}{60}\right)^4 + 4\left(\frac{1}{60}\right)^5 + 40\left(\frac{1}{60}\right)^6$$

¿Qué exactitud tenía su aproximación?

2.7 Una visión general de métodos y de software

En este capítulo hemos estudiado el problema de resolver la ecuación $f(x) = 0$ donde f es una función continua determinada. Todos los métodos comienzan con una aproximación inicial y generan una sucesión que converge a una raíz de la ecuación, si el método es exitoso. Si $[a, b]$ es un intervalo donde $f(a)$ y $f(b)$ tienen signo diferente, entonces el método de bisección y el de posición falsa convergerá. Pero la convergencia de ambos será lenta. Por lo general, se logra una convergencia más rápida usando el método de la secante o el de Newton. Ambos requieren buenas aproximaciones iniciales, el método de la secante requiere dos y una el método de Newton; por tanto, el método de bisección o el de posición falsa pueden servir como métodos iniciales en el método de la secante o en el de Newton.

El método de Müller nos dará una convergencia rápida sin una aproximación inicial muy buena. No es tan eficiente como el método de Newton; su orden de convergencia cerca de una raíz es aproximadamente $\alpha = 1.84$, en comparación con el orden cuadrático, $\alpha = 2$, del método de Newton. Pero es mejor que el método de la secante, cuyo orden es aproximadamente $\alpha = 1.62$ y tiene la ventaja adicional de aproximar raíces complejas.

La deflación generalmente se emplea con el método de Müller, una vez que se ha determinado una raíz aproximada de un polinomio. Hecha la aproximación, aplique el método de Müller o el de Newton en el polinomio original que tenga esta raíz como aproxima-

ción inicial. El procedimiento garantizará que la raíz que está siendo aproximada sea una solución de la ecuación verdadera, no de la ecuación deflacionada. Recomendamos el uso del método de Müller para obtener todas las raíces de polinomios, tanto reales como complejas. También puede utilizarse con una función continua arbitraria.

Existen otros métodos de orden superior para determinar las raíces de polinomios. Si este tema es de su interés, le aconsejamos estudiar el método de Laguerre, el cual ofrece una convergencia cúbica y además aproxima raíces complejas (véase [Ho, pp. 176-179]) donde se incluye una explicación muy completa, el método de Jenkins-Traub (véase [JT]) y el método de Brent (consúltese [Bre]).

Otro método interesante, el de Cauchy, se asemeja al de Müller; sólo que no incurre en el problema del fracaso del método de Müller, cuando $f(x_i) = f(x_{i+1}) = f(x_{i+2})$, para alguna i. Recomendamos al lector consultar [YG, secciones 4.10, 4.11 y 5.4] donde viene una explicación interesante de este método y también más detalles sobre el método de Müller.

Con una función f y una tolerancia especificadas, un programa eficiente deberá generar una aproximación a una o varias soluciones de $f(x) = 0$, cada una con un error absoluto o relativo dentro de la tolerancia; los resultados habrán de ser generados en un tiempo razonable. Si el programa no puede realizar esta tarea, por lo menos deberá dar explicaciones lógicas de por qué no se consiguió el éxito y una indicación de cómo corregir la causa del fracaso.

La subrutina ZANLY de FORTRAN de IMSL utiliza el método de Müller con deflación para aproximar varias raíces de $f(x) = 0$. La rutina ZBREN, diseñada por R. P. Brent, usa una combinación de interpolación lineal, una interpolación cuadrática inversa semejante al método de Müller y el método de bisección, y requiere que se especifique un intervalo $[a, b]$ que contenga una raíz. Las rutinas f_zeros_fcn de C y ZREAL de FORTRAN de IMSL se basan en una variante del método de Müller y aproximan los ceros de una función real f cuando sólo se tienen aproximaciones iniciales pobres. Las rutinas para determinar los ceros de polinomios son f_zeros_poly de C y ZPORC de FORTRAN, que usan el método de Jenkins-Traub para encontrar los ceros de un polinomio real; ZPLRC, que usa el método de Laguerre para determinar los ceros de un polinomio real; y las rutinas c_zeros_poly de C y ZPOCC de FORTRAN, que usan el método de Jenkins-Traub para encontrar los ceros de un polinomio complejo.

La subrutinas c05adc de C y C05ADF y C05AZF de FORTRAN de NAG usan una combinación del método de bisección, la interpolación lineal y la extrapolación para aproximar un cero real de $f(x) = 0$ en el intervalo $[a, b]$. La subrutina C05AGF es similar a C05ADF pero sólo requiere un valor inicial, en vez de un intervalo, y regresa un intervalo que contiene una raíz. Las subrutinas C05AJF y C05AXF de FORTRAN de NAG usan un método de continuación con una iteración de secante para aproximar el cero real de una función. Además, NAG proporciona las subrutinas C05AGF y C05AFF para aproximar todos los ceros de un polinomio real o complejo, respectivamente. Ambas subrutinas usan un método modificado de Laguerre para encontrar las raíces de un polinomio.

Las subrutina fzero.f de FORTRAN de netlib usa una combinación del método de bisección y el método de la secante, desarrollada por T. J. Dekker para aproximar un cero real de $f(x) = 0$ en el intervalo $[a, b]$. Requiere la especificación de un intervalo $[a, b]$ que contenga una raíz, y regresa un intervalo con un ancho menor a la tolerancia dada. La subrutina sdzro.f de FORTRAN usa una combinación para determinar un cero real de $f(x) = 0$ en un intervalo dado $[a, b]$. Las rutinas rpzero y cpzero se pueden usar para aproximar todos los ceros de un polinomio real o complejo, respectivamente. Ambas usan el método de Newton para sistemas, que estudiaremos en el capítulo 10. Todas las rutinas tienen pre-

cisión simple y doble. Estos métodos están disponibles en el sitio de netlib en Internet, http://www.netlib.org/slatec/src.

Dentro de MATLAB, la función ROOTS sirve para calcular todas las raíces de un polinomio, tanto las reales como las complejas. Para una función arbitraria, FZERO calcula una raíz cercana a una aproximación inicial especificada con determinada tolerancia.

Maple tiene el procedimiento `fsolve` para encontrar las raíces de las ecuaciones. Por ejemplo,

```
>fsolve(x^2-x-1, x);
```

revierte los números −.6180339887 y 1.618033989. También podemos especificar una variable y un intervalo para buscar. Por ejemplo,

```
>fsolve(x^2-x-1, x,1..2);
```

revierte el número 1.618033989. `fsolve` utiliza varias técnicas especializadas que se basan en la forma particular de la ecuación o sistema de ecuaciones.

Obsérvese que, a pesar de la diversidad de los métodos, los paquetes profesionales de computación tienen como fundamento principalmente los métodos y principios que expusimos en el presente capítulo. El lector deberá ser capaz de usarlos leyendo los manuales correspondientes para entender mejor los parámetros y las especificaciones de los resultados que se obtienen.

Hay tres libros clásicos en la resolución de las ecuaciones no lineales: los de Traub [Tr], de Ostrowski [Os] y de Householder [Ho]. Además, el libro de Brent [Bre] ha sido la base de muchos de los métodos de búsqueda de raíces que se utilizan actualmente.

CAPÍTULO

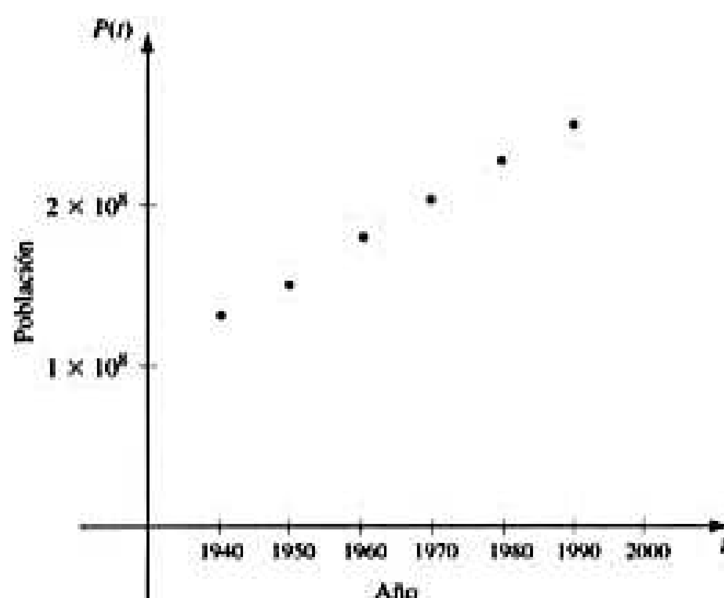
3

Interpolación y aproximación polinomial

...

Cada 10 años se levanta un censo de población en Estados Unidos. En la siguiente tabla se incluyen datos de la población, en miles de habitantes, de 1940 a 1990.

Año	1940	1950	1960	1970	1980	1990
Población en miles de habitantes	132,165	151,326	179,323	203,302	226,542	249,633



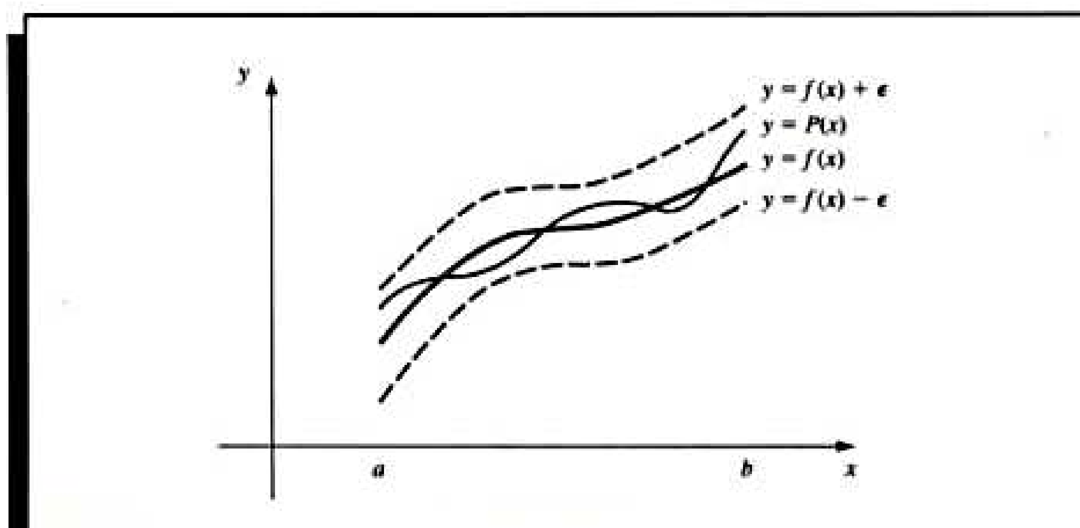
Al revisar los datos anteriores, podríamos preguntarnos si es posible utilizarlos para obtener una estimación razonable de la población que habría en —digamos— 1965 e incluso en el año 2010. Este tipo de predicciones puede obtenerse por medio de una función que corresponda a los datos disponibles. Este proceso recibe el nombre de *interpolación* y es el tema que ahora nos ocupa. Este problema demográfico se estudia a lo largo del capítulo y en los ejercicios 24 de la sección 3.1, 14 de la sección 3.2 y 24 de la sección 3.4.

Una de las clases de funciones más útiles y mejor conocidas que "manda" al conjunto de los números reales sobre sí mismo es la de los polinomios algebraicos, o sea, el conjunto de funciones de la forma

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

donde n es un entero no negativo y $a_0, \ldots, a_n$ son constantes reales. Su importancia se debe a que aproximan de manera uniforme a las funciones continuas. Dada una función cualquiera, definida y continua en un intervalo cerrado, existe un polinomio que está tan "cerca" de la función como se desee. Este resultado se expresa con precisión en el siguiente teorema. (Véase Fig. 3.1.)

Figura 3.1



Teorema 3.1 **(Teorema de aproximación de Weierstrass)**

Suponga que f está definida y es continua en $[a, b]$. Para cada $\epsilon > 0$, existe un polinomio $P(x)$, con la propiedad de que

$$|f(x) - P(x)| < \epsilon, \quad \text{para toda } x \text{ en } [a, b]. \qquad \blacksquare$$

La demostración de este teorema aparece en cualquier libro de fundamentos de análisis real (véase, por ejemplo, [Bart, pp. 165-172]).

Otra razón importante por la cual se debe considerar la clase de polinomios en la aproximación de funciones, es que la derivada y la integral indefinida de un polinomio son fá-

ciles de determinar y también son polinomios. Por estas razones, con frecuencia se usan los polinomios para aproximar a las funciones continuas.

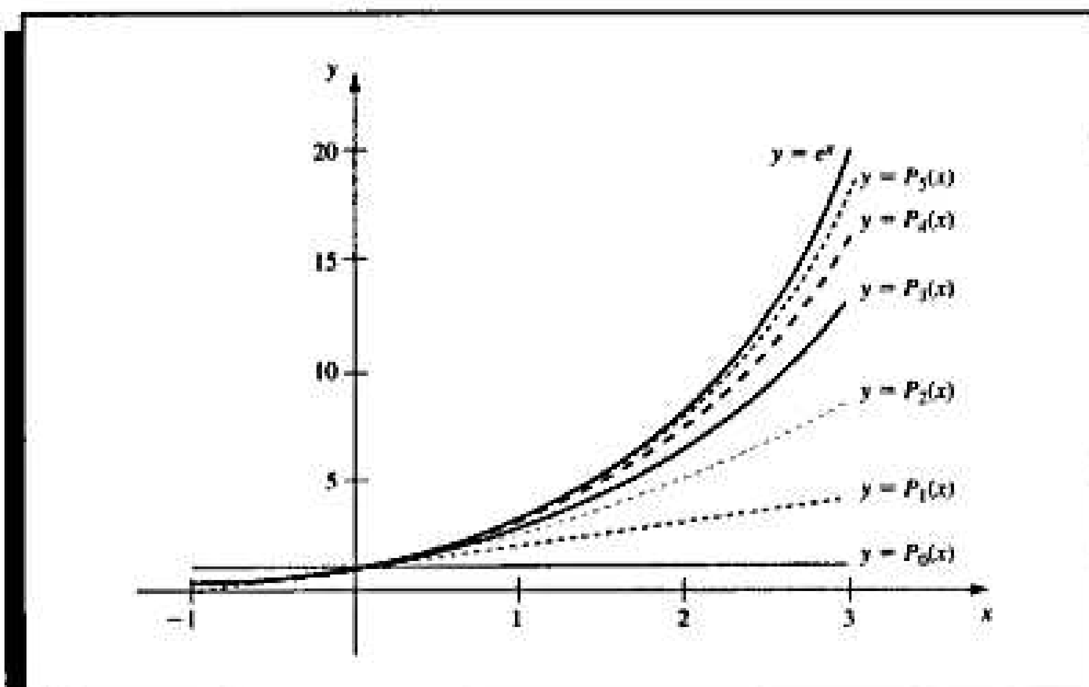
En la primera sección del libro vimos los polinomios de Taylor, y se dijo que son una de las bases fundamentales del análisis numérico. Por su importancia, cabría suponer que en la interpolación polinómica se usarían ampliamente dichas funciones, pero no es así. Los polinomios de Taylor coinciden en lo posible con determinada función en un punto específico, pero concentran su exactitud cerca de él. Una buena interpolación polinómica debe ofrecer una aproximación relativamente exacta en todo un intervalo, y los polinomios de Taylor generalmente no lo hacen. Por ejemplo, suponga que calculamos los seis primeros polinomios de Taylor alrededor de $x_0 = 0$ para $f(x) = e^x$. Como todas las derivadas de f son e^x, las cuales al evaluarse en $x_0 = 0$ dan 1, los polinomios de Taylor son

$$P_0(x) = 1, \quad P_1(x) = 1 + x, \quad P_2(x) = 1 + x + \frac{x^2}{2}, \quad P_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6},$$

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}, \quad \text{y} \quad P_5(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}.$$

Las gráficas de los polinomios se muestran en la figura 3.2. (Observe que aun en los polinomios de grado superior el error empeora progresivamente al alejarnos de cero.)

Figura 3.2



Aunque en este problema se obtienen mejores aproximaciones para $f(x) = e^x$ si utilizamos los polinomios de Taylor de grado superior, no siempre es así. Supongamos, como

un ejemplo extremo, que usamos los polinomios de diversos grados de Taylor con $f(x) = 1/x$ desarrollada alrededor de $x_0 = 1$ para aproximar $f(3) = \frac{1}{3}$. Puesto que

$$f(x) = x^{-1}, \quad f'(x) = -x^{-2}, \quad f''(x) = (-1)^2 2 \cdot x^{-3},$$

y, en general,

$$f^{(k)}(x) = (-1)^k k! x^{-k-1},$$

los polinomios de Taylor serán

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(1)}{k!}(x-1)^k = \sum_{k=0}^{n}(-1)^k (x-1)^k.$$

Si queremos aproximar $f(3) = \frac{1}{3}$ por medio de $P_n(3)$ con valores crecientes de n, obtenemos los valores en la tabla 3.1: un evidente fracaso.

Tabla 3.1

n	0	1	2	3	4	5	6	7
$P_n(3)$	1	−1	3	−5	11	−21	43	−85

El tipo de dificultad que encontramos en este caso es muy común, pues los polinomios de Taylor tienen la propiedad de que toda la información utilizada en la aproximación se concentra en el único punto x_0. Este problema generalmente limita el uso de la aproximación polinómica de Taylor al caso en que las aproximaciones se necesiten sólo en puntos cercanos a x_0. En los cálculos ordinarios conviene más usar métodos que incluyan información en diversos puntos y que estudiaremos en las siguientes páginas de este capítulo. La principal aplicación de los polinomios de Taylor en el análisis numérico no es la aproximación, sino la derivación de los métodos numéricos y la estimación del error.

3.1 Interpolación y polinomio de Lagrange

Como los polinomios de Taylor no son adecuados para la interpolación es necesario hacer uso de métodos alternos. En esta sección encontraremos polinomios de aproximación que se determinan con sólo especificar determinados puntos en el plano por donde deben pasar.

El problema de encontrar un polinomio de primer grado que pasa por los puntos distintos (x_0, y_0) y (x_1, y_1) es el mismo que el de aproximar una función f, para la cual $f(x_0) = y_0$ y $f(x_1) = y_1$ por medio de un polinomio de primer grado que interpole los valores de f en los puntos dados o que coincida con ellos. Primero definiremos las funciones

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{y} \quad L_1(x) = \frac{x - x_0}{x_1 - x_0},$$

y se define entonces

$$P(x) = L_0(x)f(x_0) + L_1(x)f(x_1).$$

Como

$$L_0(x_0) = 1, \quad L_0(x_1) = 0, \quad L_1(x_0) = 0 \quad \text{y} \quad L_1(x_1) = 1,$$

tenemos

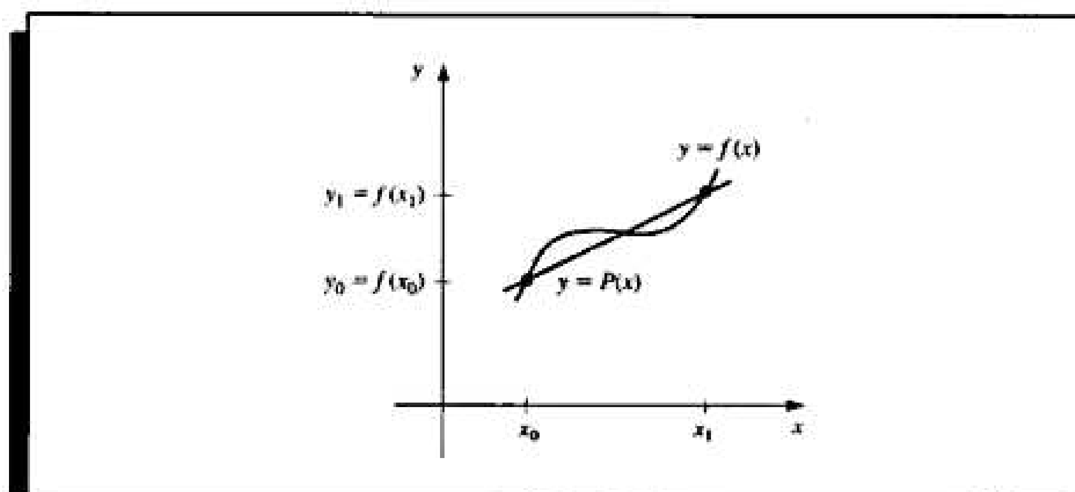
$$P(x_0) = 1 \cdot f(x_0) + 0 \cdot f(x_1) = f(x_0) = y_0$$

y

$$P(x_1) = 0 \cdot f(x_0) + 1 \cdot f(x_1) = f(x_1) = y_1.$$

Así p es la única función lineal que pasa por (x_0, y_0) y (x_1, y_1). (Véase la Fig. 3.3.)

Figura 3.3

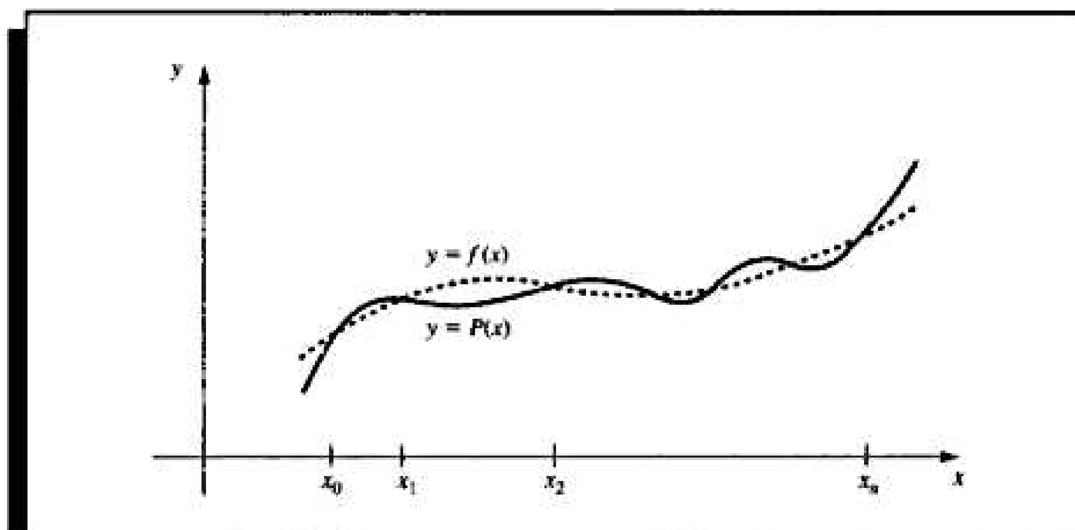


A fin de generalizar el concepto de interpolación lineal, consideremos la construcción de un polinomio de grado máximo n que pase por los $n + 1$ puntos

$$(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n)).$$

(Véase la figura 3.4.)

Figura. 3.4



En este caso para cada $k = 0, 1, \ldots, n$ construimos una función $L_{n,k}(x)$ con la propiedad de que $L_{n,k}(x_i) = 0$, cuando $i \neq k$ y $L_{n,k}(x_k) = 1$. Para satisfacer $L_{n,k}(x_i) = 0$ para cada $i \neq k$ se requiere que el numerador de $L_{n,k}(x)$ contenga el término

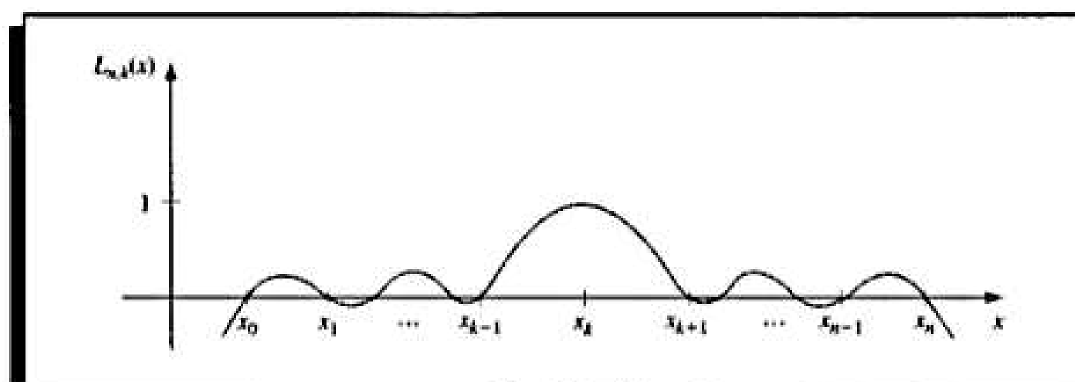
$$(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n).$$

Para satisfacer $L_{n,k}(x_k) = 1$, el denominador de $L_{n,k}(x)$ debe coincidir con este término cuando se evalúe en $x = x_k$. Es decir,

$$L_{n,k}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$

En la figura 3.5 se muestra un dibujo de la gráfica de un $L_{n,k}$ común.

Figura 3.5



El polinomio de interpolación se describe fácilmente ahora que conocemos la forma de $L_{n,k}$. Este polinomio, denominado ***n*-ésimo polinomio interpolante de Lagrange**, se define en el siguiente teorema.

Teorema 3.2 Si $x_0, x_1, \ldots, x_n$ son $n + 1$ números distintos y si f es una función cuyos valores están dados en esos números, entonces existe un único polinomio $P(x)$ de grado a lo más n, con la propiedad de que

$$f(x_k) = P(x_k) \qquad \text{para cada } k = 0, 1, \ldots, n.$$

Este polinomio está dado por

$$P(x) = f(x_0)L_{n,0}(x) + \cdots + f(x_n)L_{n,n}(x) = \sum_{k=0}^{n} f(x_k)L_{n,k}(x), \tag{3.1}$$

donde para cada $k = 0, 1, \ldots, n$,

$$L_{n,k}(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0)(x_k - x_1) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)} \tag{3.2}$$

$$= \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{(x - x_i)}{(x_k - x_i)}.$$

■

Escribiremos $L_{n,k}(x)$ simplemente como $L_k(x)$ cuando no haya confusión respecto a su grado.

EJEMPLO 1 Si queremos utilizar los números (*o nodos*) $x_0 = 2$, $x_1 = 2.5$ y $x_2 = 4$ para obtener el segundo polinomio interpolante para $f(x) = 1/x$ debemos determinar los coeficientes polinómicos $L_0(x)$, $L_1(x)$ y $L_2(x)$:

$$L_0(x) = \frac{(x - 2.5)(x - 4)}{(2 - 2.5)(2 - 4)} = (x - 6.5)x + 10,$$

$$L_1(x) = \frac{(x - 2)(x - 4)}{(2.5 - 2)(2.5 - 4)} = \frac{(-4x + 24)x - 32}{3},$$

y

$$L_2(x) = \frac{(x - 2)(x - 2.5)}{(4 - 2)(4 - 2.5)} = \frac{(x - 4.5)x + 5}{3}.$$

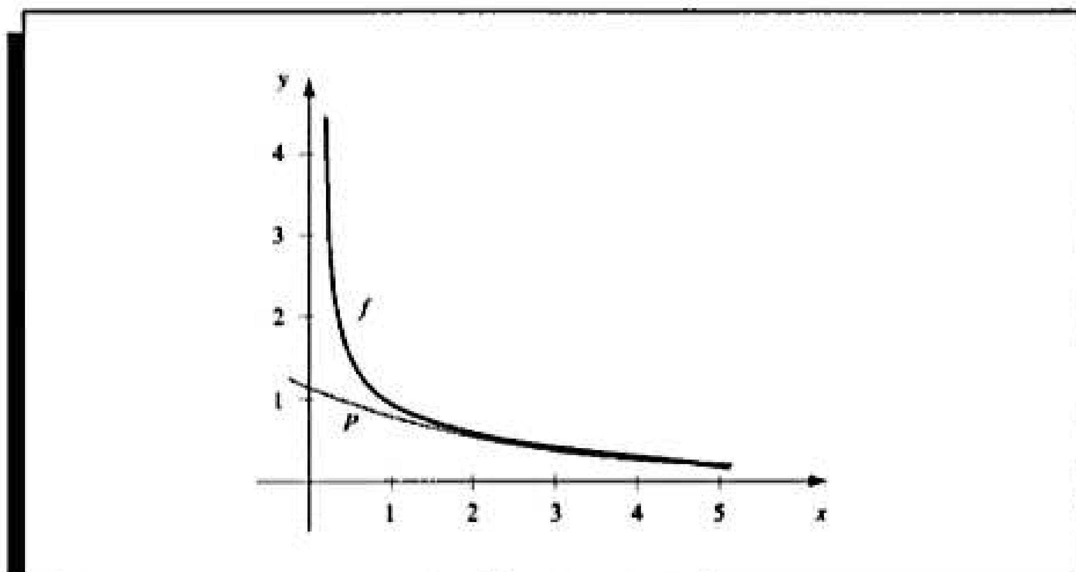
Puesto que $f(x_0) = f(2) = 0.5$, $f(x_1) = f(2.5) = 0.4$ y $f(x_2) = f(4) = 0.25$, tendremos

$$\begin{aligned} P(x) &= \sum_{k=0}^{2} f(x_k)L_k(x) \\ &= 0.5((x - 6.5)x + 10) + 0.4\,\frac{(-4x + 24)x - 32}{3} + 0.25\frac{(x - 4.5)x + 5}{3} \\ &= (0.05x - 0.425)x + 1.15. \end{aligned}$$

Una aproximación a $f(3) = \frac{1}{3}$. (Véase Fig. 3.6) es

$$f(3) \approx P(3) = 0.325.$$

Figura 3.6



Compare esto con la tabla 3.1, donde no se podía usar ningún polinomio de Taylor (desarrollado alrededor de $x_0 = 1$) para aproximar razonablemente $f(3) = \frac{1}{3}$. ■

Podemos usar un programa de cómputo para construir un polinomio interpolante. Por ejemplo, en Maple usamos

```
>interp(X, Y, x);
```

donde X es la lista $[x_0, \ldots, x_n]$, Y es la lista $[f(x_0), \ldots, f(x_n)]$ y x es la variable a ser usado. En este ejemplo podemos generar un polinomio interpolante $p = 0.05x^2 - 0.425x + 1.15$ con el comando

```
>p:=interp([2, 2.5,4], [0.5,0.4,0.25],x);
```

Para calcular $p^{(3)}$ como una aproximación a $f(3) = \frac{1}{3}$, escriba

```
>subs(x=3, p);
```

lo cual da 0.325

El siguiente paso consiste en calcular un residuo o cota del error incurrido al aproximar una función mediante un polinomio interpolante. Esto se hace en el siguiente teorema.

Teorema 3.3 Supongamos que $x_0, x_1, \ldots, x_n$ son números distintos en el intervalo $[a, b]$ y que $f \in C^{n+1}[a, b]$. Entonces, para cada x en $[a, b]$ existe un número $\xi(x)$ en (a, b) con

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n), \tag{3.3}$$

donde $P(x)$ es el polinomio interpolante de la ecuación (3.1). ■

Demostración Observe primero que, si $x = x_k$ para $k = 0, 1, \ldots, n$, entonces $f(x_k) = P(x_k)$, y al seleccionar $\xi(x_k)$ arbitrariamente en (a, b) se obtiene la ecuación (3.3). Si $x \neq x_k$ para cualquier $k = 0, 1, \ldots, n$, defina la función g para t en $[a, b]$ por medio de

$$g(t) = f(t) - P(t) - [f(x) - P(x)]\frac{(t - x_0)(t - x_1)\cdots(t - x_n)}{(x - x_0)(x - x_1)\cdots(x - x_n)}$$

$$= f(t) - P(t) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(t - x_i)}{(x - x_i)}.$$

Puesto que $f \in C^{n+1}[a, b]$, y $P \in C^{\infty}[a, b]$, se deduce que $g \in C^{n+1}[a, b]$. Cuando $t = x_k$ tendremos

$$g(x_k) = f(x_k) - P(x_k) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(x_k - x_i)}{(x - x_i)} = 0 - [f(x) - P(x)]\cdot 0 = 0.$$

Además,

$$g(x) = f(x) - P(x) - [f(x) - P(x)]\prod_{i=0}^{n}\frac{(x - x_i)}{(x - x_i)} = f(x) - P(x) - [f(x) - P(x)] = 0.$$

Por tanto, $g \in C^{n+1}[a, b]$, y g se anula en los $n + 2$ números distintos $x, x_0, x_1, \ldots, x_n$. Conforme al teorema generalizado de Rolle, existe ξ en (a, b) tal que $g^{(n+1)}(\xi) = 0$.

Por tanto,

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P^{(n+1)}(\xi) - [f(x) - P(x)] \frac{d^{n+1}}{dt^{n+1}} \left[\prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)} \right]_{t=\xi} \tag{3.4}$$

Por ser $P(x)$ un polinomio de grado a lo más n, su $(n + 1)$-ésima derivada, $P^{(n+1)}(x)$, será igual a cero. Asimismo, $\prod_{i=0}^{n} [(t - x_i)/(x - x_i)]$ es un polinomio de grado $(n + 1)$ y, por tanto,

$$\prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)} = \left[\frac{1}{\prod_{i=0}^{n} (x - x_i)} \right] t^{n+1} + (\text{término de menor grado en } t),$$

y

$$\frac{d^{n+1}}{dt^{n+1}} \prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)} = \frac{(n + 1)!}{\prod_{i=0}^{n} (x - x_i)}.$$

La ecuación (3.4) ahora se convierte en

$$0 = f^{(n+1)}(\xi) - 0 - [f(x) - P(x)] \frac{(n + 1)!}{\prod_{i=0}^{n} (x - x_i)},$$

y luego de despejar $f(x)$, tendremos

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi)}{(n + 1)!} \prod_{i=0}^{n} (x - x_i).$$

■ ■ ■

La fórmula de error obtenida en el teorema 3.3 es un resultado teórico muy importante, porque los polinomios de Lagrange se emplean frecuentemente para deducir la diferenciación numérica y los métodos de integración. Las cotas de error de estas técnicas se obtienen aplicando la fórmula del error de Lagrange.

Nótese que la forma del error del polinomio de Lagrange se parece mucho a la del polinomio de Taylor. El polinomio de Taylor de grado n alrededor de x_0 concentra en x_0 toda la información conocida y tiene un término de error de la forma

$$\frac{f^{(n+1)}(\xi(x))}{(n + 1)!} (x - x_0)^{n+1}.$$

El polinomio de Lagrange de grado n utiliza información en los números distintos x_0, $x_1, \ldots, x_n$ y, en lugar de $(x - x_0)^n$, su fórmula de error utiliza un producto de $n + 1$ términos $(x - x_0), (x - x_1), \ldots, (x - x_n)$:

$$\frac{f^{(n+1)}(\xi(x))}{(n + 1)!} (x - x_0)(x - x_1) \cdots (x - x_n).$$

El uso específico de esta fórmula de error se limita a las funciones cuyas derivadas tienen cotas conocidas.

EJEMPLO 2 Suponga que debe preparar una tabla de la función $f(x) = e^x$, para x en $[0, 1]$. Suponga, además, que el número de cifras decimales de cada entrada o valor es $d \geq 8$ y que h, el tamaño del paso, es la diferencia entre los valores adyacentes de x. ¿Cuál debe ser el valor de h para que la interpolación lineal (es decir, el polinomio de grado 1 de Lagrange) arroje un error absoluto a lo máximo de 10^{-6}?

Sean $x_0, x_1, \ldots,$ los números en los que se evalúa f, en x en $[0, 1]$, y suponga que j satisface $x_j \le x \le x_{j+1}$. La ecuación (3.3) significa que el error de la interpolación lineal es

$$|f(x) - P(x)| = \left|\frac{f^{(2)}(\xi)}{2!}(x - x_j)(x - x_{j+1})\right| = \frac{|f^{(2)}(\xi)|}{2}|(x - x_j)||(x - x_{j+1})|.$$

Por ser h el tamaño del paso, se deduce que $x_j = jh$, $x_{j+1} = (j + 1)h$, y que

$$|f(x) - P(x)| = \frac{|f^{(2)}(\xi)|}{2!}|(x - jh)(x - (j + 1)h)|.$$

Por tanto,

$$|f(x) - P(x)| \le \frac{1}{2}\max_{\xi \in [0,1]} e^{\xi} \max_{x_j \le x \le x_{j+1}} |(x - jh)(x - (j + 1)h)|$$

$$\le \frac{1}{2}e \max_{x_j \le x \le x_{j+1}} |(x - jh)(x - (j + 1)h)|.$$

Al considerar $g(x) = (x - jh)(x - (j + l)h)$ para $jh \le x \le (j + 1)h$ y al aplicar los métodos de cálculo (véase Ejercicio 28), encontramos que

$$\max_{x_j \le x \le x_{j+1}} |(x - jh)(x - (j + 1)h)| = \max_{x_j \le x \le x_{j+1}} |g(x)| = \left|g\left(\left(j + \frac{1}{2}\right)h\right)\right| = \frac{h^2}{4}.$$

En consecuencia, el error de la interpolación lineal está acotado por

$$|f(x) - P(x)| \le \frac{eh^2}{8},$$

y es suficiente elegir h de modo que

$$\frac{eh^2}{8} \le 10^{-6}, \text{ lo cual implica que } h < 1.72 \times 10^{-3}.$$

Puesto que $n = (1 - 0)/h$ debe ser un entero, una elección lógica del tamaño del paso es $h = 0.001$. ■

En el siguiente ejemplo se explica la interpolación cuando no es posible emplear la parte de la ecuación (3.3) correspondiente al error.

EJEMPLO 3 La tabla 3.2 muestra los valores de una función en diversos puntos. Compararemos las aproximaciones a $f(1.5)$ obtenidas con varios polinomios de Lagrange.

Tabla 3.2

x	$f(x)$
1.0	0.7651977
1.3	0.6200860
1.6	0.4554022
1.9	0.2818186
2.2	0.1103623

Como 1.5 se halla entre 1.3 y 1.6, el polinomio lineal utilizará $x_0 = 1.3$ y $x_1 = 1.6$. El valor del polinomio interpolante en 1.5 es

$$P_1(1.5) = \frac{(1.5-1.6)}{(1.3-1.6)}(0.6200860) + \frac{(1.5-1.3)}{(1.6-1.3)}(0.4554022) = 0.5102968.$$

Es razonable emplear dos polinomios de grado 2: uno suponiendo que $x_0 = 1.3$, $x_1 = 1.6$ y que $x_2 = 1.9$, lo cual nos da

$$P_2(1.5) = \frac{(1.5-1.6)(1.5-1.9)}{(1.3-1.6)(1.3-1.9)}(0.6200860) + \frac{(1.5-1.3)(1.5-1.9)}{(1.6-1.3)(1.6-1.9)}(0.4554022)$$

$$+ \frac{(1.5-1.3)(1.5-1.6)}{(1.9-1.3)(1.9-1.6)}(0.2818186)$$

$$= 0.5112857,$$

y el otro suponiendo que $x_0 = 1.0$, $x_1 = 1.3$, y que $x_2 = 1.6$, lo cual nos da

$$\hat{P}_2(1.5) = 0.5124715.$$

En el caso del tercer grado hay dos formas de elegir el polinomio. Una consiste en suponer que $x_0 = 1.3$, $x_1 = 1.6$, $x_2 = 1.9$ y que $x_3 = 2.2$, lo cual nos da

$$P_3(1.5) = 0.5118302.$$

La otra consiste en suponer que $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, y que $x_3 = 1.9$, lo cual nos da

$$\hat{P}_3(1.5) = 0.5118127.$$

El polinomio de Lagrange de cuarto grado utiliza todas las entradas o valores de la tabla. Cuando $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$ y cuando $x_4 = 2.2$, la aproximación es

$$P_4(1.5) = 0.5118200.$$

Esperamos obtener este grado de exactitud con las aproximaciones anteriores, ya que $P_3(1.5)$, $\hat{P}_3(1.5)$ y $P_4(1.5)$ coinciden con una exactitud de 2×10^{-5} unidades. También esperamos que $P_4(1.5)$ sea la aproximación más exacta, porque emplea una mayor cantidad de los datos proporcionados.

La función que estamos aproximando es la función de Bessel de primer tipo de orden cero, cuyo valor en 1.5 es de 0.5118277; por tanto, éstas son las verdaderas exactitudes de las aproximaciones:

$$|P_1(1.5) - f(1.5)| \approx 1.53 \times 10^{-3},$$
$$|P_2(1.5) - f(1.5)| \approx 5.42 \times 10^{-4},$$
$$|\hat{P}_2(1.5) - f(1.5)| \approx 6.44 \times 10^{-4},$$
$$|P_3(1.5) - f(1.5)| \approx 2.5 \times 10^{-6},$$
$$|\hat{P}_3(1.5) - f(1.5)| \approx 1.50 \times 10^{-5},$$
$$|P_4(1.5) - f(1.5)| \approx 7.7 \times 10^{-6},$$

Adviértase que $P_3(1.5)$ es la aproximación más exacta; pero si no conocemos el valor real de $f(1.5)$ aceptaríamos $P_4(1.5)$ como la mejor aproximación, ya que utiliza una mayor cantidad de los datos proporcionados. En este caso no podemos servirnos del término del error o término residual derivado en el teorema 3.3, ya que no conocemos la cuarta derivada de f. Desafortunadamente, casi siempre ocurre esto. ∎

Una dificultad práctica que ocurre con la interpolación de Lagrange consiste en que el término del error es difícil de aplicar, generalmente el grado del polinomio necesario para lograr la exactitud deseada no se conoce antes de determinar los cálculos. Se acostumbra obtener los resultados a partir de varios polinomios, hasta que se logra una correspondencia apropiada como en el ejemplo anterior. Además, el trabajo realizado al calcular la aproximación mediante el segundo polinomio no reduce el que se requiere para calcular el tercero; tampoco es más fácil obtener la cuarta aproximación, una vez conocida la tercera y así sucesivamente. A continuación derivaremos estos polinomios de aproximación de tal forma que se aprovechen mejor los cálculos anteriores.

Definición 3.4 Sea f una función definida en $x_0, x_1, x_2, \ldots, x_n$, y supongamos que $m_1, m_2, \ldots, m_k$, son k enteros distintos con $0 \leq m_i \leq n$ para cada i. El polinomio de Lagrange que concuerda con f en los k puntos $x_{m_1}, x_{m_2}, \ldots, x_{m_k}$ se denota por $P_{m_1, m_2, \ldots, m_k}(x)$. ∎

EJEMPLO 4 Si $x_0 = 1, x_1 = 2, x_2 = 3, x_3 = 4, x_4 = 6$, y si $f(x) = e^x$, entonces $P_{1,2,4}(x)$ será el polinomio que concuerda con $f(x)$ en $x_1 = 2$, $x_2 = 3$ y con $x_4 = 6$, es decir,

$$P_{1,2,4}(x) = \frac{(x-3)(x-6)}{(2-3)(2-6)}e^2 + \frac{(x-2)(x-6)}{(3-2)(3-6)}e^3 + \frac{(x-2)(x-3)}{(6-2)(6-3)}e^6.$$

∎

En el siguiente resultado se describe un método con el que se generan recursivamente aproximaciones al polinomio de Lagrange.

Teorema 3.5 Si f está definida en $x_0, x_1, \ldots, x_k$, y x_j y x_i son dos números distintos de este conjunto, entonces

$$P(x) = \frac{(x - x_j)P_{0,1,\ldots,j-1,j+1,\ldots,k}(x) - (x - x_i)P_{0,1,\ldots,i-1,i+1,\ldots,k}(x)}{(x_i - x_j)}$$

describe el polinomio de grado k de Lagrange que interpola f en los $k + 1$ puntos $x_0, x_1, \ldots, x_k$. ∎

Demostración Para facilitar la notación, sean $Q \equiv P_{0,1,\ldots,i-1,i+1,\ldots,k}$ y $\hat{Q} \equiv P_{0,1,\ldots,j-1,j+1,\ldots,k}$. Puesto que $Q(x)$ y $\hat{Q}(x)$ son polinomios de grado $k - 1$ o menos, $P(x)$ será de grado a lo más k. Si $0 \leq r \leq k$ y si $r \neq i, j$, entonces $Q(x_r) = \hat{Q}(x_r) = f(x_r)$, así que

$$P(x_r) = \frac{(x_r - x_j)\hat{Q}(x_r) - (x_r - x_i)Q(x_r)}{x_i - x_j} = \frac{(x_i - x_j)}{(x_i - x_j)}f(x_r) = f(x_r).$$

Además, como $\hat{Q}(x_i) = f(x_i)$, tenemos

$$P(x_i) = \frac{(x_i - x_j)\hat{Q}(x_i) - (x_i - x_i)Q(x_i)}{x_i - x_j} = \frac{(x_i - x_j)}{(x_i - x_j)}f(x_i) = f(x_i).$$

De modo análogo, como $Q(x_j) = f(x_j)$, obtenemos $P(x_j) = f(x_j)$. Pero, por definición, $P_{0,1,\ldots,k}(x)$ es el polinomio único de grado a lo más k que concuerda con f en $x_0, x_1, \ldots, x_k$. Por consiguiente, $P \equiv P_{0,1,\ldots,k}$. ■ ■ ■

De acuerdo con el teorema 3.5, los polinomios interpolantes pueden generarse de manera recursiva. Por ejemplo, podemos generarlos como se indica en la tabla 3.3, donde cada hilera se termina antes de iniciar las siguientes.

Tabla 3.3

x_0	$P_0 = Q_{0,0}$				
x_1	$P_1 = Q_{1,0}$	$P_{0,1} = Q_{1,1}$			
x_2	$P_2 = Q_{2,0}$	$P_{1,2} = Q_{2,1}$	$P_{0,1,2} = Q_{2,2}$		
x_3	$P_3 = Q_{3,0}$	$P_{2,3} = Q_{3,1}$	$P_{1,2,3} = Q_{3,2}$	$P_{0,1,2,3} = Q_{3,3}$	
x_4	$P_4 = Q_{4,0}$	$P_{3,4} = Q_{4,1}$	$P_{2,3,4} = Q_{4,2}$	$P_{1,2,3,4} = Q_{4,3}$	$P_{0,1,2,3,4} = Q_{4,4}$

A este procedimiento se le conoce con el nombre de **método de Neville**. La notación P que se usa en la tabla 3.3 es difícil de manejar por la cantidad de subíndices con que se representan las entradas o datos. Pero obsérvese que, al construir un arreglo, sólo se necesitan dos subíndices. Descender por la tabla equivale a utilizar puntos consecutivos x_i con i más grande; desplazarse hacia la derecha equivale a aumentar el grado del polinomio interpolante. Dado que los puntos aparecen consecutivamente en cada entrada, debemos describir únicamente un punto inicial y la cantidad de puntos adicionales con que se construirá la aproximación.

Para evitar los subíndices múltiples, sea $Q_{i,j}(x)$, $0 \leq j \leq i$, el polinomio interpolante de grado j en los $(j + 1)$ números $x_{i-j}, x_{i-j+1}, \ldots, x_{i-1}, x_i$; es decir,

$$Q_{i,j} = P_{i-j,\, i-j+1, \ldots,\, i-1,\, i}.$$

Al aplicar esta notación en el método de Neville se obtiene el arreglo de la notación Q de la tabla 3.3.

EJEMPLO 5 En el ejemplo 3, los valores de diversos polinomios interpolantes en $x = 1.5$ se obtuvieron por medio de los datos incluidos en las dos primeras columnas de la tabla 3.4. En este ejemplo, aproximamos $f(1.5)$ usando el resultado del teorema 3.5 Si $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$ y si $x_4 = 2.2$, entonces $Q_{0,0} = f(1.0)$, $Q_{1,0} = f(1.3)$, $Q_{2,0} = f(1.6)$, $Q_{3,0} = f(1.9)$, y $Q_{4,0} = f(2.2)$. Estos son los cinco polinomios de grado cero (constante) que aproximan $f(1.5)$.

Al calcular las aproximaciones de primer grado $Q_{1,1}(1.5)$, obtendremos

$$\begin{aligned} Q_{1,1}(1.5) &= \frac{(x - x_0)Q_{1,0} - (x - x_1)Q_{0,0}}{x_1 - x_0} \\ &= \frac{(1.5 - 1.0)Q_{1,0} - (1.5 - 1.3)Q_{0,0}}{1.3 - 1.0} \\ &= \frac{0.5(0.6200860) - 0.2(0.7651977)}{0.3} = 0.5233449. \end{aligned}$$

De manera análoga,

$$Q_{2,1}(1.5) = \frac{(1.5 - 1.3)(0.4554022) - (1.5 - 1.6)(0.600860)}{1.6 - 1.3} = 0.5102968,$$

$$Q_{3,1}(1.5) = 0.5132634 \quad \text{y} \quad Q_{4,1}(1.5) = 0.5104270.$$

Se espera que la mejor aproximación lineal sea $Q_{2,1}$, ya que 1.5 se encuentra entre $x_1 = 1.3$ y $x_2 = 1.6$.

En forma parecida, las aproximaciones usando los polinomios de grado superior están dadas por

$$Q_{2,2}(1.5) = \frac{(1.5 - 1.0)(0.5102968) - (1.5 - 1.6)(0.5233449)}{1.6 - 1.0} = 0.5124715,$$

$$Q_{3,2}(1.5) = 0.5112857 \qquad \text{y} \qquad Q_{4,2}(1.5) = 0.5137361.$$

Las aproximaciones de grado superior se generan de modo semejante y se incluyen en la tabla 3.4. ■

Tabla 3.4

1.0	0.7651977				
1.3	0.6200860	0.5233449			
1.6	0.4554022	0.5102968	0.5124715		
1.9	0.2818186	0.5132634	0.5112857	0.5118127	
2.2	0.1103623	0.5104270	0.5137361	0.5118302	0.5118200

Si la última aproximación, $Q_{4,4}$, no ofrece la exactitud deseada, podemos seleccionar otro nodo, x_5 y agregar otra hilera o renglón a la tabla:

$$x_5 \quad Q_{5,0} \quad Q_{5,1} \quad Q_{5,2} \quad Q_{5,3} \quad Q_{5,4} \quad Q_{5,5}.$$

Entonces podemos comparar $Q_{4,4}$, $Q_{5,4}$ y $Q_{5,5}$ para tener aún más exactitud.

En el ejemplo, a esta función se le conoce como de Bessel de primer tipo de orden cero, cuyo valor en 2.5 es −0.0483838. Con esto podemos construir una nueva hilera o renglón de aproximaciones a $f(1.5)$:

$$2.5 \quad -0.0483838 \quad 0.4807699 \quad 0.5301984 \quad 0.5119070 \quad 0.5118430 \quad 0.5118277.$$

La última entrada o valor, 0.5118277, es correcta a siete cifras decimales.

EJEMPLO 6 La tabla 3.5 contiene los valores de $f(x) = \ln x$ con una precisión de cifras decimales dada.

Tabla 3.5

i	x_i	$\ln x_i$
0	2.0	0.6931
1	2.2	0.7885
2	2.3	0.8329

Nos serviremos del método de Neville para aproximar $f(2.1) = \ln 2.1$. Al completar la tabla, da los valores.

Tabla 3.6

i	x_i	$x-x_i$	Q_{i0}	Q_{i1}	Q_{i2}
0	2.0	0.1	0.6931		
1	2.2	−0.1	0.7885	0.7410	
2	2.3	−0.2	0.8329	0.7441	0.7420

Por tanto, $P_2(2.1) = Q_{22} = 0.7420$. Puesto que $f(2.1) = \ln 2.1 = 0.7419$ con cuatro lugares decimales de exactitud, el error absoluto será

$$\begin{aligned}|f(2.1) - P_2(2.1)| &= |0.7419 - 0.7420| \\ &= 10^{-4}.\end{aligned}$$

Sin embargo, $f'(x) = 1/x$, $f''(x) = -1/x^2$, y $f'''(x) = 2/x^3$, así que la fórmula (3.3) da una cota de error

$$\begin{aligned}|f(2.1) - P_2(2.1)| &= \left|\frac{f'''(\xi)}{3!}(x - x_0)(x - x_1)(x - x_2)\right| \\ &= \left|\frac{1}{3\xi^3}(0.1)(-0.1)(-0.2)\right| \leq 8.\bar{3} \times 10^{-5}.\end{aligned}$$

Nótese que el error real, 10^{-4}, rebasa la cota de error $8.\bar{3} \times 10^{-5}$. Esta contradicción aparente es consecuencia de los cálculos con un número finito de dígitos. Hemos usado las aproximaciones de cuatro dígitos, y la fórmula del error (3.3) supone una aritmética de dígitos infinitos. A ello se debe que nuestros errores reales sean mayores que la estimación teórica. ■

En el algoritmo 3.1 se construyen por renglones las entradas o datos del método de Neville.

ALGORITMO 3.1

Interpolación iterada de Neville

Para evaluar el polinomio interpolante P en los $n + 1$ números distintos $x_0, \ldots, x_n$ en el número x para la función f:

ENTRADA los números $x, x_0, x_1, \ldots, x_n$; valores $f(x_0), f(x_1), \ldots, f(x_n)$ como la primera columna $Q_{0,0}, Q_{1,0}, \ldots, Q_{n,0}$ de Q.

SALIDA la tabla Q con $P(x) = Q_{n,n}$.

Paso 1 Para $i = 1, 2, \ldots, n$
para $j = 1, 2, \ldots, i$

$$\text{tome } Q_{i,j} = \frac{(x - x_{i-j})Q_{i,j-1} - (x - x_i)Q_{i-1,j-1}}{x_i - x_{i-j}}.$$

Paso 2 SALIDA (Q);
PARAR. ■

Se puede modificar el algoritmo para agregar nuevos nodos interpolantes. Por ejemplo, podemos hacer uso de la desigualdad

$$|Q_{i,i} - Q_{i-1,i-1}| < \varepsilon$$

como criterio de paro, donde ε es una tolerancia prescrita del error. Si la desigualdad es verdadera, $Q_{i,i}$ será una aproximación razonable a $f(x)$. Si es falsa, se agrega un nuevo punto de interpolación x_{i+1}.

CONJUNTO DE EJERCICIOS 3.1

1. Para las funciones dadas $f(x)$, sean $x_0 = 0$, $x_1 = 0.6$ y $x_2 = 0.9$. Construya polinomios de interpolación de grados uno y dos a lo máximo para aproximar $f(0.45)$, y calcule el error real.
   a. $f(x) = \cos x$
   b. $f(x) = \sqrt{1+x}$
   c. $f(x) = \ln(x+1)$
   d. $f(x) = \tan x$
2. Aplique el teorema 3.3 para calcular la cota de error en las aproximaciones del ejercicio 1.
3. Use los polinomios interpolantes de Lagrange apropiados de grados uno, dos y tres para aproximar lo siguiente:
   a. $f(8.4)$ si $f(8.1) = 16.94410$, $f(8.3) = 17.56492$, $f(8.6) = 18.50515$, $f(8.7) = 18.82091$
   b. $f\left(-\frac{1}{3}\right)$ si $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.334993750$, $f(0) = 1.10100000$
   c. $f(0.25)$ si $f(0.1) = 0.62049958$, $f(0.2) = -0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$
   d. $f(0.9)$ si $f(0.6) = -0.17694460$, $f(0.7) = 0.01375227$, $f(0.8) = 0.22363362$, $f(1.0) = 0.65809197$
4. Aplique el método de Neville para obtener las aproximaciones del ejercicio 3.
5. Aplique el método de Neville para aproximar $\sqrt{3}$ con la función $f(x) = 3^x$ y los valores $x_0 = -2$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, y $x_4 = 2$.
6. Aplique el método de Neville para aproximar $\sqrt{3}$ con la función $f(x) = \sqrt{x}$ y los valores $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $x_3 = 4$ y $x_4 = 5$. Compare la exactitud con la del ejercicio 5.
7. Los datos del ejercicio 3 se generaron usando las siguientes funciones. Use la fórmula correspondiente para encontrar una cota de error y compare la cota con el error real cuando $n = 1$ y $n = 2$.
   a. $f(x) = x \ln x$
   b. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$
   c. $f(x) = x \cos x - 2x^2 + 3x - 1$
   d. $f(x) = \operatorname{sen}(e^x - 2)$
8. Sean $f(x) = \sqrt{x - x^2}$ y $P_2(x)$ el polinomio interpolante en $x_0 = 0$, x_1 y $x_2 = 1$. Calcule el valor más grande de x_1 en $(0, 1)$ para el cual $f(0.5) - P_2(0.5) = -0.25$.
9. Sea $P_3(x)$ el polinomio interpolante para los datos $(0, 0)$, $(0.5, y)$, $(1, 3)$ y $(2, 2)$. Encuentre y si el coeficiente de x^3 en $P_3(x)$ es 6.
10. Use el polinomio interpolante de Lagrange de grado tres o menos y la aritmética de corte a cuatro dígitos para aproximar 0.750 por medio de los siguientes valores. Calcule una cota de error para la aproximación.

    $\cos 0.698 = 0.7661$ $\quad \cos 0.733 = 0.7432$ $\quad \cos 0.768 = 0.7193$ $\quad \cos 0.803 = 0.6946$

    El valor real de cos 0.750 es 0.7317 (con una exactitud de cuatro cifras decimales). Explique la discrepancia existente entre el error real y la cota de error.
11. Use los siguientes valores y la aritmética de redondeo a cuatro dígitos para construir una aproximación del tercer polinomio de Lagrange a $f(1.09)$. La función que va a ser aproximada es $f(x) = \log_{10}(\tan x)$. Conociendo lo anterior, calcule una cota del error en la aproximación.

    $f(1.00) = 0.1924$ $\quad f(1.05) = 0.2414$ $\quad f(1.10) = 0.2933$ $\quad f(1.15) = 0.3492$

12. Repita el ejercicio 11 usando Maple y la aritmética de redondeo a diez dígitos.
13. El método de Neville sirve para aproximar $f(0.5)$, cuando se dispone de la siguiente tabla.

$x_0 = 0$	$P_0 = 0$		
$x_1 = 0.4$	$P_1 = 2.8$	$P_{01} = 3.5$	
$x_2 = 0.7$	P_2	P_{12}	$P_{012} = \frac{27}{7}$

Determine $P_2 = f(0.7)$.
14. El método de Neville sirve para aproximar $f(0.4)$, se cuenta con la siguiente tabla.

$x_0 = 0$	$P_0 = 1$			
$x_1 = 0.25$	$P_1 = 2$	$P_{01} = 2.6$		
$x_2 = 0.5$	P_2	P_{12}	P_{012}	
$x_3 = 0.75$	$P_3 = 8$	$P_{23} = 2.4$	$P_{123} = 2.96$	$P_{0123} = 3.016$

Determine $P_2 = f(0.5)$.
15. Construya los polinomios interpolantes de Lagrange para las siguientes funciones y obtenga una cota del error absoluto en el intervalo $[x_0, x_n]$.
   a. $f(x) = e^{2x} \cos 3x$, $x_0 = 0, x_1 = 0.3, x_2 = 0.6, n = 2$
   b. $f(x) = \text{sen}(\ln x)$, $x_0 = 2.0, x_1 = 2.4, x_2 = 2.6, n = 2$
   c. $f(x) = \ln x$, $x_0 = 1, x_1 = 1.1, x_2 = 1.3, x_3 = 1.4, n = 3$
   d. $f(x) = \cos x + \text{sen}\, x$, $x_0 = 0, x_1 = 0.25, x_2 = 0.5, x_3 = 1.0, n = 3$
16. Sea $f(x) = e^x$, para $0 \leq x \leq 2$.
   a. Aproxime $f(0.25)$ mediante la interpolación lineal con $x_0 = 0$ y $x_1 = 0.5$.
   b. Aproxime $f(0.75)$ mediante la interpolación lineal con $x_0 = 0.5$ y $x_1 = 1$.
   c. Aproxime $f(0.25)$ y $f(0.75)$ mediante el segundo polinomio interpolante con $x_0 = 0, x_1 = 1$ y $x_2 = 2$.
   d. ¿Cuáles aproximaciones son mejores y por qué?
17. Suponga que necesita construir tablas de ocho lugares decimales para la función logarítmica común, o de base 10, de $x = 1$ a $x = 10$, de modo que la interpolación lineal tenga una exactitud de 10^{-6}. Determine una cota del tamaño del paso para esta tabla. ¿Qué tamaño de paso escogerá para asegurarse de que la tabla incluya $x = 10$?
18. Suponga que $x_j = j$ para $j = 0, 1, 2, 3$ y que se sabe que

$$P_{0,1}(x) = x + 1, \qquad P_{1,2}(x) = 3x - 1 \qquad \text{y} \qquad P_{1,2,3}(1.5) = 4.$$

Obtenga $P_{0,1,2,3}(1.5)$.
19. Suponga que $x_j = j$ para $j = 0, 1, 2$ y 3 y que se sabe que

$$P_{0,1}(x) = 2x + 1, \qquad P_{0,2}(x) = x + 1 \qquad \text{y} \qquad P_{1,2,3}(2.5) = 3.$$

Obtenga $P_{0,1,2,3}(2.5)$.
20. El algoritmo de Neville sirve para aproximar $f(0)$ por medio de $f(-2), f(-1), f(1)$ y $f(2)$. Suponga que $f(-1)$ se sobreexpresó por 2 y que $f(1)$ se subexpresó en 3. Determine el error del cálculo original del valor del polinomio interpolante al aproximar $f(0)$.
21. Construya una sucesión de valores interpolantes y_n a $f(1 + \sqrt{10})$ donde $f(x) = (1 + x^2)^{-1}$ para $-5 \leq x \leq 5$, como sigue: para cada $n = 1, 2, \ldots, 10$, sea $h = 10/n$ y $y_n = P_n(1 + \sqrt{10})$,

donde $P_n(x)$ es el polinomio interpolante de $f(x)$ en los nodos $x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}$ y $x_j^{(n)} = -5 + jh$ para cada $j = 0, 1, 2, \ldots, n$. ¿Parece que la sucesión $\{y_n\}$ converge en $f(1 + \sqrt{10})$?

Interpolación inversa Suponga que $f \in C^1[a, b]$, $f'(x) \neq 0$ en $[a, b]$ y que f tiene un cero p en $[a, b]$. Sean $x_0, \ldots, x_n$, $n + 1$ números distintos en $[a, b]$ con $f(x_k) = y_k$ para cada $k = 0, 1, \ldots, n$. Si quiere aproximar p, construya el polinomio interpolante de grado n en los nodos $y_0, \ldots, y_n$ para f^{-1}. Puesto que $y_k = f(x_k)$ y $0 = f(p)$, se deduce que $f^{-1}(y_k) = x_k$ y $p = f^{-1}(0)$. Se da el nombre de *interpolación iterada inversa* al uso de la interpolación iterada para aproximar $f^{-1}(0)$.

22. Use la interpolación iterada inversa para obtener una aproximación a la solución de $x - e^{-x} = 0$, por medio de los datos

x	0.3	0.4	0.5	0.6
e^{-x}	0.740818	0.670320	0.606531	0.548812

23. Construya un algoritmo que sirva para obtener la interpolación inversa.

24. **a.** En la introducción de este capítulo se incluyó una tabla con la población de Estados Unidos, entre los años 1940 y 1990. Use la interpolación de Lagrange para aproximar la población en los años, 1930, 1965 y 2010.

b. La población de 1930 fue aproximadamente de 123 203 000 habitantes. ¿Qué exactitud, a su juicio, tienen sus cifras correspondientes a los años 1965 y 2010?

25. Se sospecha que las elevadas concentraciones de tanina en las hojas de los robles maduros inhiben el crecimiento de las larvas de la polilla invernal (*Operophtera bromata L., Geometridae*) que tanto dañan a los árboles en algunos años. La tabla anexa contiene el peso promedio de dos muestras de larva, tomadas en los primeros 28 días después del nacimiento. La primera muestra se crió en hojas de robles jóvenes, mientras que la segunda lo hizo en hojas maduras del mismo árbol.

a. Use la interpolación de Lagrange para aproximar la curva del peso promedio de las muestras.

b. Para calcular un peso promedio máximo aproximado de cada muestra, determine el máximo del polinomio interpolante.

Día	0	6	10	13	17	20	28
Peso promedio de la muestra 1 (mg)	6.67	17.33	42.67	37.33	30.10	29.31	28.74
Peso promedio de la muestra 2 (mg)	6.67	16.11	18.89	15.00	10.56	9.44	8.89

26. En el ejercicio 24 de la sección 1.1 se integró una serie de Maclaurin para aproximar erf(1), donde erf(x) es la distribución normal de la función de error definida por

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

a. Use la serie de Maclaurin para construir una tabla de erf(x) con una exactitud de 10^{-4} para erf(x_i) donde $x_i = 0.2i$, para $i = 0, 1, \ldots, 5$.

b. Use la interpolación lineal y la cuadrática para obtener una aproximación de erf$(\frac{1}{3})$ ¿Qué método le parece más adecuado?

27. Demuestre el teorema 1.14 aplicando el procedimiento de la demostración del teorema 3.3. [*Sugerencia*: sea

$$g(t) = f(t) - P(t) - [f(x) - P(x)] \cdot \frac{(t - x_0)^{n+1}}{(x - x_0)^{n+1}},$$

donde P es el polinomio de Taylor de n-ésimo grado y use el teorema 1.12.]

28. Demuestre que $\max\limits_{x_j \le x \le x_{j+1}} |g(x)| = h^2/4$, donde $g(x) = (x - jh)(x - (j + 1)h)$.

29. El polinomio de Bernstein de grado n para $f \in C[0, 1]$ está dado por

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1 - x)^{n-k},$$

donde $\binom{n}{k}$ denota $n!/k!(n - k)!$. Estos polinomios pueden usarse en una demostración constructiva del teorema de aproximación de Weierstrass 3.1 (véase [Bart]), ya que $\lim\limits_{n\to\infty} B_n(x) = f(x)$, para cada $x \in [0,1]$.

a. Obtenga $B_3(x)$ para las funciones

(i) $f(x) = x$ (ii) $f(x) = 1$

b. Demuestre que, para cada $k \le n$,

$$\binom{n-1}{k-1} = \left(\frac{k}{n}\right)\binom{n}{k}.$$

c. Utilice la parte (b) y el hecho de que, según (ii) de la parte (a),

$$1 = \sum_{k=0}^{n} \binom{n}{k} x^k (1 - x)^{n-k}, \qquad \text{para cada } n,$$

para demostrar que cuando $f(x) = x^2$,

$$B_n(x) = \left(\frac{n-1}{n}\right) x^2 + \frac{1}{n} x.$$

d. Utilice la parte (c) para estimar el valor de n necesario para que $|B_n(x) - x^2| \le 10^{-6}$ sea válido para todas las x en $[0, 1]$.

3.2 Diferencias divididas

En la sección anterior utilizamos la interpolación iterada para generar aproximaciones polinómicas de grado cada vez mayor en un punto específico. Los métodos de diferencias divididas, que explicaremos en esta sección, sirven para generar sucesivamente los polinomios. El estudio que haremos de este tema será breve, pues los resultados de esta sección no tendrán gran uso en lo que resta del libro. En la mayor parte de los textos antiguos de análisis numéricos se examinan de modo exhaustivo los métodos de las diferencias divididas. Si usted necesita un tratamiento más completo, le recomendamos consultar el libro de Hildebrand [Hild].

Supongamos que $P_n(x)$ es el n-ésimo polinomio de Lagrange que concuerda con la función f en los números distintos $x_0, x_1, \ldots, x_n$. Las diferencias divididas de f respecto a $x_0, x_1, \ldots, x_n$ se usan para expresar $P_n(x)$ en la forma

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}), \tag{3.5}$$

para las constantes apropiadas $a_0, a_1, \ldots, a_n$.

Para determinar la primera de las constantes, a_0 note que, si $P_n(x)$ está escrito en la forma de la ecuación (3.5), entonces al evaluar $P_n(x)$ en x_0 queda sólo el término constante a_0; es decir

$$a_0 = P_n(x_0) = f(x_0).$$

De manera similar, cuando se evalúa $P(x)$ en x_1, los únicos términos no cero en la evaluación de $P_n(x_1)$ son los términos constante y lineal,

$$f(x_0) + a_1(x_1 - x_0) = P_n(x_1) = f(x_1);$$

así que

$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \tag{3.6}$$

Ahora es necesario presentar la notación de diferencias divididas, que nos recuerda la notación Δ^2 de Aitken que utilizamos en la sección 2.5. **La diferencia dividida cero** de la función f respecto a x_i, que se denota como $f[x_i]$, es simplemente el valor de f en x_i:

$$f[x_i] = f(x_i). \tag{3.7}$$

El resto de las diferencias divididas se definen en forma inductiva. La **primera diferencia dividida** de f respecto a x_i y x_{i+1} se denota $f[x_i, x_{i+1}]$ y se define así

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}. \tag{3.8}$$

La **segunda diferencia dividida** $f[x_i, x_{i+1}, x_{i+2}]$ se define como sigue

$$f[x_i, x_{i+1}, x_{x+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}.$$

En forma análoga, después de determinar las primeras $(k - 1)$ diferencias divididas,

$$f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}] \qquad \text{y} \qquad f[x_{i+1}, x_{i+2}, \ldots, x_{i+k-1}, x_{i+k}],$$

la ***k*-ésima diferencia dividida** relativa a $x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+k}$ está dada por

$$f[x_i, x_{i+1}, \ldots, x_{i+k-1}, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}. \tag{3.9}$$

Con esta notación, podemos reexpresar la ecuación (3.6) como $a_1 = f[x_0, x_1]$ y el polinomio interpolante de la ecuación (3.5) es

$$\begin{aligned} P_n(x) = f[x_0] &+ f[x_0, x_1](x - x_0) + a_2(x - x_0)(x - x_1) \\ &+ \cdots + a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

Como cabe suponer tras evaluar a_0 y a_1, las constantes requeridas son

$$a_k = f[x_0, x_1, x_2, \ldots, x_k],$$

para cada $k = 0, 1, \ldots, n$. Por tanto, podemos reescribir $P_n(x)$ como (véase [Hild, pp. 43-47)]

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0) \cdots (x - x_{k-1}). \tag{3.10}$$

Como se indica en el ejercicio 17, el valor de $f[x_0, x_1, ..., x_k]$ es independiente del orden de los números $x_0, x_1, ..., x_k$. A esta ecuación se le conoce con el nombre de **fórmula de diferencias divididas interpolantes de Newton.** En la tabla 3.7 se describe esquemáticamente la determinación de las diferencias divididas obtenida de los puntos de datos tabulados. Con esos datos también es posible determinar dos cuartas diferencias y una quinta diferencia.

Tabla 3.7

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas
x_0	$f[x_0]$			
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$	
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{x_2 - x_1}$		$f[x_0, x_1, x_2, x_3] = \dfrac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$
x_2	$f[x_2]$		$f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$	
		$f[x_2, x_3] = \dfrac{f[x_3] - f[x_2]}{x_3 - x_2}$		$f[x_1, x_2, x_3, x_4] = \dfrac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1}$
x_3	$f[x_3]$		$f[x_2, x_3, x_4] = \dfrac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2}$	
		$f[x_3, x_4] = \dfrac{f[x_4] - f[x_3]}{x_4 - x_3}$		$f[x_2, x_3, x_4, x_5] = \dfrac{f[x_3, x_4, x_5] - f[x_2, x_3, x_4]}{x_5 - x_2}$
x_4	$f[x_4]$		$f[x_3, x_4, x_5] = \dfrac{f[x_4, x_5] - f[x_3, x_4]}{x_5 - x_3}$	
		$f[x_4, x_5] = \dfrac{f[x_5] - f[x_4]}{x_5 - x_4}$		
x_5	$f[x_5]$			

La fórmula de las diferencias divididas interpolantes de Newton puede implantarse por medio del algoritmo 3.2. Se puede modificar la forma de la salida para producir todas las diferencias divididas, como se hizo en el ejemplo 1.

ALGORITMO 3.2

Fórmula de las diferencias divididas interpolantes de Newton

Para obtener los coeficientes de las diferencias divididas del polinomio interpolante P en los $(n + 1)$ números distintos $x_0, x_1, \ldots, x_n$; para la función f:

ENTRADA los números $x_0, x_1, \ldots, x_n$; valores $f(x_0), f(x_1), \ldots, f(x_n)$ como $F_{0,0}, F_{1,0}, \ldots, F_{n,0}$.

SALIDA los números $F_{0,0}, F_{1,1}, \ldots, F_{n,n}$ donde

$$P(x) = \sum_{i=0}^{n} F_{i,i} \prod_{j=0}^{i-1} (x - x_j).$$

Paso 1 Para $i = 1, 2, \ldots, n$
para $j = 1, 2, \ldots, i$
tome $F_{i,j} = \dfrac{F_{i,j-1} - F_{i-1,j-1}}{x_i - x_{i-j}}$.

Paso 2 SALIDA $(F_{0,0}, F_{1,1}, \ldots, F_{n,n})$; ($F_{i,i}$ es $f[x_0, x_1, \ldots, x_i]$.)
PARAR. ■

EJEMPLO 1 En el ejemplo 3 de la sección 3.1 utilizamos varios polinomios interpolantes para aproximar $f(1.5)$, por medio de los datos contenidos en las tres primeras columnas de la tabla 3.8. El resto de entradas o datos de la tabla incluyen diferencias divididas que se calcularon mediante el algoritmo 3.2.

Los coeficientes de la fórmula de las diferencias divididas progresivas del polinomio interpolante de Newton se encuentran a lo largo de la diagonal de la tabla. El polinomio es

$$\begin{aligned} P_4(x) = &\ 0.7651977 - 0.4837057\,(x - 1.0) - 0.1087339(x - 1.0)(x - 1.3) \\ &+ 0.0658784(x - 1.0)(x - 1.3)(x - 1.6) \\ &+ 0.0018251(x - 1.0)(x - 1.3)(x - 1.6)(x - 1.9). \end{aligned}$$

Nótese que el valor $P_4(1.5) = 0.5118200$ concuerda con el resultado de la sección 3.1, ejemplo 3, como debe ser, porque los polinomios son los mismos. ■

Tabla 3.8

i	x_i	$f[x_i]$	$f[x_{i-1}, x_i]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, \ldots, x_i]$	$f[x_{i-4}, \ldots, x_i]$
0	1.0	0.7651977				
			−0.4837057			
1	1.3	0.6200860		−0.1087339		
			−0.5489460		0.0658784	
2	1.6	0.4554022		−0.0494433		0.0018251
			−0.5786120		0.0680685	
3	1.9	0.2818186		0.0118183		
			−0.5715210			
4	2.2	0.1103623				

El teorema del valor medio aplicado a la ecuación (3.8) cuando $i = 0$,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

significa que, cuando existe f', $f[x_0, x_1] = f'(\xi)$ para algún número ξ entre x_0 y x_1. El siguiente teorema generaliza este resultado.

Teorema 3.6 Supongamos que $f \in C^n[a, b]$ y $x_0, x_1, \ldots, x_n$ son números distintos en $[a, b]$. Entonces existe un número ξ en (a, b) con,

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

■

Demostración Sea

$$g(x) = f(x) - P_n(x).$$

Puesto que $f(x_i) = P_n(x_i)$ para cada $i = 0, 1, \ldots, n$, la función g tiene $n + 1$ ceros distintos en $[a, b]$. Conforme al teorema generalizado de Rolle, existe en (a, b) un número ξ con $g^{(n)}(\xi) = 0$, tal que

$$0 = f^{(n)}(\xi) - P_n^{(n)}(\xi).$$

Por ser $P_n(x)$ un polinomio de grado n cuyo coeficiente principal es $f[x_0, x_1, \ldots, x_n]$,

$$P_n^{(n)}(x) = n!\, f[x_0, x_1, \ldots, x_n],$$

para todos los valores de x. En consecuencia,

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

■ ■ ■

La fórmula de las diferencias divididas interpolantes de Newton puede expresarse en forma simplificada cuando se arreglan consecutivamente $x_0, x_1, \ldots, x_n$ con espacios iguales. Al introducir la notación $h = x_{i+1} - x_i$ para cada $i = 0, 1, \ldots, n - 1$ y sea $x = x_0 + sh$, podemos escribir la diferencia $x - x_i$ como $x - x_i = (s - i)h$. Por tanto, la ecuación (3.10) se transforma en

$$\begin{aligned} P_n(x) = P_n(x_0 + sh) &= f[x_0] + shf[x_0, x_1] + s(s-1)h^2 f[x_0, x_1, x_2] \\ &+ \cdots + s(s-1)(s-n+1)h^n f[x_0, x_1, \ldots, x_n] \\ &= \sum_{k=0}^{n} s(s-1)\cdots(s-k+1)h^k f[x_0, x_1, \ldots, x_k]. \end{aligned}$$

Al utilizar la notación del coeficiente binomial,

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

podemos expresar $P_n(x)$ en forma compacta como

$$P_n(x) = P_n(x_0 + sh) = f[x_0] + \sum_{k=1}^{n} \binom{s}{k} k!h^k f[x_0, x_1, \ldots, x_k]. \qquad (3.11)$$

A ésta se le llama **fórmula de las diferencias divididas progresivas de Newton**. Otra forma, denominada **fórmula de las diferencias progresivas de Newton**, se construye utilizando la notación de las diferencias progresivas Δ que explicamos al hablar del método Δ^2 de Aitken. Con esta notación,

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{1}{h}\Delta f(x_0)$$

$$f[x_0, x_1, x_2] = \frac{1}{2h}\left[\frac{\Delta f(x_1) - \Delta f(x_0)}{h}\right] = \frac{1}{2h^2}\Delta^2 f(x_0),$$

y, en general,

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!h^k}\,\Delta^k f(x_0).$$

Entonces, la ecuación (3.11) tiene la siguiente fórmula.

Fórmula de las diferencias progresivas de Newton

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} \binom{s}{k} \Delta^k f(x_0) \tag{3.12}$$

Si reordenamos los nodos interpolantes como $x_n, x_{n-1}, \ldots, x_0$ se obtiene una fórmula semejante a la ecuación (3.10):

$$\begin{aligned} P_n(x) &= f[x_n] + f[x_n, x_{n-1}](x - x_n) + f[x_n, x_{n-1}, x_{n-2}](x - x_n)(x - x_{n-1}) \\ &\quad + \cdots + f[x_n, \ldots, x_0](x - x_n)(x - x_{n-1}) \cdots (x - x_1). \end{aligned}$$

Si los nodos tienen espacios iguales con $x = x_n + sh$ y $x = x_i + (s + n - i)h$ entonces

$$\begin{aligned} P_n(x) &= P_n(x_n + sh) \\ &= f[x_n] + sh\,f[x_n, x_{n-1}] + s(s+1)h^2 f[x_n, x_{n-1}, x_{n-2}] + \cdots \\ &\quad + s(s+1)\cdots(s+n-1)h^n f[x_n, \ldots, x_0]. \end{aligned}$$

Esta forma se conoce con el nombre de **fórmula de las diferencias divididas regresivas de Newton**, y sirve para derivar una fórmula de uso más común denominada **fórmula de las diferencias regresivas de Newton**. Para explicar esta última, necesitamos la siguiente definición.

Definición 3.7 Dada la sucesión $\{p_n\}_{n=0}^{\infty}$, defina la diferencia regresiva ∇p_n (léase *nabla* p_n) por medio de

$$\nabla p_n = p_n - p_{n-1}, \qquad \text{para } n \geq 1.$$

Las potencias mayores se definen recursivamente por

$$\nabla^k p_n = \nabla(\nabla^{k-1} p_n), \qquad \text{para } k \geq 2. \quad \blacksquare$$

La definición 3.7 implica que

$$f[x_n, x_{n-1}] = \frac{1}{h}\nabla f(x_n), \qquad f[x_n, x_{n-1}, x_{n-2}] = \frac{1}{2h^2}\,\nabla^2 f(x_n),$$

y, en general,

$$f[x_n, x_{n-1}, \ldots, x_{n-k}] = \frac{1}{k!h^k}\,\nabla^k f(x_n).$$

En consecuencia,

$$P_n(x) = f[x_n] + s\nabla f(x_n) + \frac{s(s+1)}{2}\,\nabla^2 f(x_n) + \cdots + \frac{s(s+1)\cdots(s+n-1)}{n!}\,\nabla^n f(x_n).$$

La notación del coeficiente binomial se amplió, para incluir todos los valores reales de s al tomar

$$\binom{-s}{k} = \frac{-s(-s-1)\cdots(-s-k+1)}{k!} = (-1)^k \frac{s(s+1)\cdots(s+k-1)}{k!},$$

por tanto,

$$P_n(x) = f[x_n] + (-1)^1\binom{-s}{1}\nabla f(x_n) + (-1)^2\binom{-s}{2}\nabla^2 f(x_n) + \cdots + (-1)^n\binom{-s}{n}\nabla^n f(x_n).$$

Esto nos da el siguiente resultado.

Fórmula de las diferencias regresivas de Newton

$$P_n(x) = f[x_n] + \sum_{k=1}^{n} (-1)^k \binom{-s}{k} \nabla^k f(x_n) \tag{3.13}$$

EJEMPLO 2 La tabla 3.9 corresponde a las diferencias divididas de los datos que se dan en el ejemplo 1.

Tabla 3.9

		Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	Cuartas diferencias divididas
1.0	0.7651977				
		−0.4837057			
1.3	0.6200860		−0.1087339		
		−0.5489460		0.0658784	
1.6	0.4554022		−0.0494433		0.0018251
		−0.5786120		0.0680685	
1.9	0.2818186		0.0118183		
		−0.5715210			
2.2	0.1103623				

Sólo un polinomio interpolante hasta de grado 4 usa estos cinco puntos, pero organizaremos los puntos para obtener mejores aproximaciones de interpolación de grados 1, 2 y 3. Esto nos dará la exactitud de una aproximación de cuarto grado para el valor dado de x.

Si se requiere una aproximación a $f(1.1)$ una elección adecuada de los nodos será $x_0 = 1.0$, $x_1 = 1.3$, $x_2 = 1.6$, $x_3 = 1.9$ y $x_4 = 2.2$, porque es la que utiliza lo antes posible los puntos de datos más cercanos a $x = 1.1$, y también hace uso de la cuarta diferencia dividida. Ello significa que $h = 0.3$ y que $s = \frac{1}{3}$, por lo cual la fórmula se emplea con las diferencias divididas que aparecen subrayadas con líneas seguidas en la tabla 3.9:

$$P_4(1.1) = P_4(1.0 + \frac{1}{3}(0.3))$$

$$= 0.7651997 + \frac{1}{3}(0.3)(-0.4837057) + \frac{1}{3}\left(-\frac{2}{3}\right)(0.3)^2(-0.1087339)$$

$$+ \frac{1}{3}\left(-\frac{2}{3}\right)\left(-\frac{5}{3}\right)(0.3)^3(0.0658784)$$
$$+ \frac{1}{3}\left(-\frac{2}{3}\right)\left(-\frac{5}{3}\right)\left(-\frac{8}{3}\right)(0.3)^4(0.0018251)$$
$$= 0.7196480.$$

Si queremos aproximar un valor cuando x está cerca del final de los valores tabulados, digamos $x = 2.0$, de nuevo sería conveniente utilizar lo antes posible los puntos de datos más cercanos a x. Para ello es necesario aplicar la fórmula de diferencias divididas regresivas con $s = -\frac{2}{3}$ y las diferencias divididas de la tabla 3.9 que aparecen subrayadas con líneas *punteadas*:

$$P_4(2.0) = P_4\left(2.2 - \frac{2}{3}(0.3)\right)$$
$$= 0.1103623 - \frac{2}{3}(0.3)(-0.5715210) - \frac{2}{3}\left(\frac{1}{3}\right)(0.3)^2(0.0118183)$$
$$- \frac{2}{3}\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)(0.3)^3(0.0680685) - \frac{2}{3}\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)\left(\frac{7}{3}\right)(0.3)^4(0.0018251)$$
$$= 0.2238754.$$ ■

Las fórmulas de Newton no son convenientes para aproximar un valor $f(x)$ para x situado cerca del centro de la tabla, porque x_0 no podrá estar cerca de x si empleamos el método regresivo o el progresivo, de modo que intervenga la diferencia de orden más alto. En este caso disponemos de varias fórmulas de diferencias divididas; cada una de ellas es aplicable de manera óptima en determinadas situaciones. A esas técnicas se les llama **fórmulas de diferencias centradas**. Hay varias de ellas, pero por ahora sólo nos ocuparemos de una, el método de Stirling; una vez más, al lector que desee una explicación más completa le aconsejamos consultar a Hildebrand [Hild].

Para las fórmulas de diferencias centradas escogemos x_0 cerca del punto que va a ser aproximado y marcamos los nodos directamente por debajo de x_0, como $x_1, x_2, \ldots$ y como los que están directamente arriba como $x_{-1}, x_{-2}, \ldots$. Con esta convención, la **fórmula de Stirling** está dada por

$$P_n(x) = P_{2m+1}(x) = f[x_0] + \frac{sh}{2}(f[x_{-1}, x_0] + f[x_0, x_1]) + s^2h^2 f[x_{-1}, x_0, x_1] \tag{3.14}$$
$$+ \frac{s(s^2-1)h^3}{2} f[x_{-2}, x_{-1}, x_0, x_1] + f[x_{-1}, x_0, x_1, x_2])$$
$$+ \cdots + s^2(s^2-1)(s^2-4)\cdots(s^2-(m-1)^2)h^{2m} f[x_{-m}, \ldots, x_m]$$
$$+ \frac{s(s^2-1)\cdots(s^2-m^2)h^{2m+1}}{2}(f[x_{-m-1}, \ldots, x_m] + f[x_{-m}, \ldots, x_{m+1}]),$$

si $n = 2m + 1$ es impar, y si $n = 2m$ es par, aplicamos la misma fórmula pero suprimimos la última línea. Los elementos de esta fórmula aparecen subrayados en la tabla 3.10.

EJEMPLO 3 Considere la tabla de datos que se dio en los ejemplos precedentes. Si queremos aplicar la fórmula de Stirling para aproximar $f(1.5)$ con $x_0 = 1.6$, usamos los elementos *subrayados* en la tabla de diferencias 3.11.

Tabla 3.10

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	Cuartas diferencias divididas
x_{-2}	$f[x_{-2}]$				
		$f[x_{-2}, x_{-1}]$			
x_{-1}	$f[x_{-1}]$		$f[x_{-2}, x_{-1}, x_0]$		
		$\underline{f[x_{-1}, x_0]}$		$\underline{f[x_{-2}, x_{-1}, x_0, x_1]}$	
x_0	$\underline{f[x_0]}$		$\underline{f[x_{-1}, x_0, x_1]}$		$\underline{f[x_{-2}, x_{-1}, x_0, x_1, x_2]}$
		$\underline{f[x_0, x_1]}$		$\underline{f[x_{-1}, x_0, x_1, x_2]}$	
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$			
x_2	$f[x_2]$				

Tabla 3.11

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	Cuartas diferencias divididas
1.0	0.7651977				
		−0.4837057			
1.3	0.6200860		−0.1087339		
		<u>−0.5489460</u>		<u>0.0658784</u>	
1.6	<u>0.4554022</u>		<u>−0.0494433</u>		<u>0.0018251</u>
		<u>−0.5786120</u>		<u>0.0680685</u>	
1.9	0.2818186		0.0118183		
		−0.5715210			
2.2	0.1103623				

La fórmula con $h = 0.3$, $x_0 = 1.6$ y $s = -\frac{1}{3}$, se convierte en

$$
\begin{aligned}
f(1.5) &\approx P_4\left(1.6 + \left(-\frac{1}{3}\right)(0.3)\right) \\
&= 0.4554022 + \left(-\frac{1}{3}\right)\left(\frac{0.3}{2}\right)((-0.5489460) + (-0.5786120)) \\
&+ \left(-\frac{1}{3}\right)^2 (0.3)^2(-0.0494433) \\
&+ \frac{1}{2}\left(-\frac{1}{3}\right)\left(\left(-\frac{1}{3}\right)^2 - 1\right)(0.3)^3\,(0.0658784 + 0.0680685) \\
&+ \left(-\frac{1}{3}\right)^2\left(\left(-\frac{1}{3}\right)^2 - 1\right)(0.3)^4(0.0018251) \\
&= 0.5118200.
\end{aligned}
$$

■

CONJUNTO DE EJERCICIOS 3.2

1. Use la fórmula de diferencias divididas interpolantes de Newton o el algoritmo 3.2 para construir polinomios interpolantes de grado uno, dos y tres con los siguientes datos. Use cada uno de los polinomios para aproximar el valor especificado.

   a. $f(8.4)$ si $f(8.1) = 16.94410$, $f(8.3) = 17.56492$, $f(8.6) = 18.50515$, $f(8.7) = 18.82091$

   b. $f(0.9)$ si $f(0.6) = -0.17694460$, $f(0.7) = 0.01375227$, $f(0.8) = 0.22363362$, $f(1.0) = 0.65809197$

2. Use la fórmula de diferencia progresiva de Newton para construir polinomios interpolantes de grado uno, dos y tres con los siguientes datos. Aproxime el valor especificado usando cada uno de los polinomios.

   a. $f(-\frac{1}{3})$ si $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.33493750$, $f(0) = 1.10100000$

   b. $f(0.25)$ si $f(0.1) = -0.62049958$, $f(0.2) = -0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$

3. Use la fórmula de diferencias regresivas de Newton para construir polinomios interpolantes de grado uno, dos y tres con los siguientes datos. Por medio de cada uno de los polinomios aproxime el valor especificado.

   a. $f(-\frac{1}{3})$ si $f(-0.75) = -0.07181250$, $f(-0.5) = -0.02475000$, $f(-0.25) = 0.33493750$, $f(0) = 1.10100000$

   b. $f(0.25)$ si $f(0.1) = -0.62049958$, $f(0.2) = 0.28398668$, $f(0.3) = 0.00660095$, $f(0.4) = 0.24842440$

4. a. Use el algoritmo 3.2 para construir el polinomio interpolante de grado cuatro con los puntos desigualmente espaciados que aparecen en la tabla anexa:

x	$f(x)$
0.0	−6.00000
0.1	−5.89483
0.3	−5.65014
0.6	−5.17788
1.0	−4.28172

   b. Agregue $f(1.1) = -3.99583$ a la tabla y construya el polinomio interpolante de grado cinco.

5. a. Aproxime $f(0.05)$ mediante los siguientes datos y la fórmula de diferencias divididas progresivas de Newton:

x	0.0	0.2	0.4	0.6	0.8
$f(x)$	1.00000	1.22140	1.49182	1.82212	2.22554

   b. Use la fórmula de las diferencias divididas regresivas de Newton para aproximar $f(0.65)$.

   c. Aplique la fórmula de Stirling para aproximar $f(0.43)$.

6. Demuestre que el polinomio que interpola los siguientes datos es de grado 3.

x	−2	−1	0	1	2	3
$f(x)$	1	4	11	16	13	−4

7. a. Demuestre que los polinomios de diferencias divididas progresivas de Newton

$$P(x) = 3 - 2(x + 1) + 0(x + 1)(x) + (x + 1)(x)(x - 1)$$

y

$$Q(x) = -1 + 4(x + 2) - 3(x + 2)(x + 1) + (x + 2)(x + 1)(x)$$

interpolan los datos

x	-2	-1	0	1	2
$f(x)$	-1	3	1	-1	3

b. ¿Por qué la parte (a) no viola la propiedad de singularidad de los polinomios interpolantes?

8. Un polinomio de cuarto grado $P(x)$ satisface $\Delta^4 P(0) = 24$, $\Delta^3 P(0) = 6$ y $\Delta^2 P(0) = 0$, donde $\Delta P(x) = P(x + 1) - P(x)$. Calcule $\Delta^2 P(10)$.

9. Se tienen los siguientes datos para un polinomio $P(x)$ de grado desconocido.

x	0	1	2
$P(x)$	2	-1	4

Determine el coeficiente de x^2 en $P(x)$ si todas las diferencias progresivas de tercer orden son 1.

10. Se dan los siguientes datos para un polinomio $P(x)$ de grado desconocido.

x	0	1	2	3
$P(x)$	4	9	15	18

Determine el coeficiente de x^3 en $P(x)$ si todas las diferencias progresivas de cuarto orden son 1.

11. La fórmula de las diferencias divididas progresivas de Newton sirve para aproximar $f(0.3)$ si se cuenta con los siguientes datos.

x	0.0	0.2	0.4	0.6
$f(x)$	15.0	21.0	30.0	51.0

Suponga que se descubre que $f(0.4)$ fue subexpresado en 10 y que $f(0.6)$ fue sobreexpresado en 5. ¿Cuánto deberá modificarse la aproximación a $f(0.3)$?

12. Con una función f la fórmula de las diferencias divididas interpolantes de Newton da el polinomio interpolante

$$P_3(x) = 1 + 4x + 4x(x - 0.25) + \frac{16}{3}x(x - 0.25)(x - 0.5),$$

en los nodos $x_0 = 0$, $x_1 = 0.25$, $x_2 = 0.5$ y $x_3 = 0.75$. Obtenga $f(0.75)$.

13. Con una función f las diferencias divididas progresivas están dadas por

$x_0 = 0.0$	$f[x_0]$		
		$f[x_0, x_1]$	
$x_1 = 0.4$	$f[x_1]$		$f[x_0, x_1, x_2] = \frac{50}{7}$
		$f[x_1, x_2] = 10$	
$x_2 = 0.7$	$f[x_2] = 6$		

Determine los datos que faltan en la tabla.

14. a. En la introducción de este capítulo se incluyó una tabla que contiene la población de Estados Unidos de 1940 a 1990. Utilice las diferencias divididas adecuadas para aproximar la población de los años 1930, 1965 y 2010.
    b. En 1930 la población fue aproximadamente de 123 203 000. ¿Cuál es, a su juicio, la exactitud de las cifras correspondientes a los años de 1965 y 2010?
15. Si se tiene

$$\begin{aligned}P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + a_2(x - x_0)(x - x_1)\\ + a_3(x - x_0)(x - x_1)(x - x_2) + \cdots\\ + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}),\end{aligned}$$

use $P_n(x_2)$ para demostrar que $a_2 = f[x_0, x_1, x_2]$.

16. Demuestre que

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi(x))}{(n+1)!},$$

para alguna $\xi(x)$ [*Sugerencia*: según la ecuación (3.3),

$$f(x) = P_n(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)\cdots(x - x_n).$$

Si consideramos el polinomio interpolante de grado $n + 1$ en $x_0, x_1, \ldots, x_n, x$, tenemos

$$f(x) = P_{n+1}(x) = P_n(x) + f[x_0, x_1, \ldots, x_n, x]\,(x - x_0)\cdots(x - x_n).]$$

17. Sea $i_0, i_1, \ldots, i_n$ un rearreglo de los enteros $0, 1, \ldots, n$. Demuestre que $f[x_{i_0}, x_{i_1}, \ldots, x_{i_n}] = f[x_0, x_1, \ldots, x_n]$. [*Sugerencia*: considere el coeficiente principal del polinomio de Lagrange de grado n en los datos $\{x_0, x_1, \ldots, x_n\} = \{x_{i_0}, x_{i_1}, \ldots, x_{i_n}\}$.]

3.3 Interpolación de Hermite

Los *polinomios osculantes* representan una generalización de los polinomios de Taylor y de Lagrange. Dados $n + 1$ números distintos $x_0, x_1, \ldots, x_n$ en $[a, b]$ y los enteros no negativos $m_0, m_1, \ldots, m_n$, y $m = \text{máx}\{m_0, m_1, \ldots, m_n\}$. El polinomio osculante que aproxima una función $f \in C^m[a, b]$, en x_i, para cada $i = 0, \ldots, n$, es el polinomio de menor grado que concuerda con la función f y con todas sus derivadas de orden menor o igual que m_i en x_i para cada $i = 0, 1, \ldots, n$. El grado de este polinomio osculante es, a lo más,

$$M = \sum_{i=0}^{n} m_i + n$$

ya que el número de condiciones por cumplir es $\sum_{i=0}^{n} m_i + (n + 1)$, y un polinomio de grado M tiene $M + 1$ coeficientes que podemos utilizar para satisfacerlas.

Definición 3.8 Sean $x_0, x_1, \ldots, x_n$, $n + 1$ números distintos en $[a, b]$ y m_i un entero no negativo asociado a x_i para $i = 0, 1, \ldots, n$. Supóngase que $f \in C^m[a, b]$ y que $m = \text{máx}_{0 \le i \le n} m_i$. El **polinomio osculante** que aproxima f es el polinomio $P(x)$ de menor grado tal que

$$\frac{d^k P(x_i)}{dx^k} = \frac{d^k f(x_i)}{dx^k} \quad \text{para cada } i = 0, 1, \ldots, n \quad \text{y} \quad k = 0, 1, \ldots, m_i. \quad \blacksquare$$

Nótese que, cuando $n = 0$, el polinomio osculante que aproxima f es simplemente el polinomio m_0-ésimo de Taylor para f en x_0. Cuando $m_i = 0$ para cada i, el polinomio osculante es el n-ésimo polinomio de Lagrange que interpola f en $x_0, x_1, \ldots, x_n$.

Cuando $m_i = 1$ para cada $i = 0, 1, \ldots, n$, se produce una clase de polinomios denominados **polinomios de Hermite**. En una función dada f, estos últimos concuerdan con f en $x_0, x_1, \ldots, x_n$. Además, como sus primeras derivadas concuerdan con las de f, tendrán la misma "forma" que la función en $(x_i, f(x_i))$, en el sentido de que las *líneas tangentes* del polinomio coinciden con las de la función. Aquí estudiaremos sólo los polinomios osculantes en esta situación y examinaremos primero un teorema que describe con precisión la forma de los polinomios de Hermite.

Teorema 3.9 Si $f \in C^1[a, b]$ y si $x_0, \ldots, x_n \in [a, b]$ son distintos, el polinomio único de menor grado que concuerda con f y f' en $x_0, \ldots, x_n$ es el polinomio de Hermite de grado a lo más $2n + 1$ que está dado por

$$H_{2n+1}(x) = \sum_{j=0}^{n} f(x_j)\, H_{n,j}(x) + \sum_{j=0}^{n} f'(x_j)\, \hat{H}_{n,j}(x),$$

donde

$$H_{n,j}(x) = [1 - 2(x - x_j)L'_{n,j}(x_j)]L^2_{n,j}(x)$$

y

$$\hat{H}_{n,j}(x) = (x - x_j)L^2_{n,j}(x).$$

Dentro de este contexto $L_{n,j}(x)$ denota el j-ésimo polinomio de Lagrange de grado n definido en la ecuación (3.2).

Más aún, si $f \in C^{2n+2}[a, b]$ entonces para $x \in [a, b]$

$$f(x) = H_{2n+1}(x) + \frac{(x - x_0)^2 \cdots (x - x_n)^2}{(2n + 2)!} f^{(2n+2)}(\xi),$$

para alguna ξ con $a < \xi < b$. $\blacksquare$

Demostración Ante todo, recuerde que

$$L_{n,j}(x_i) = \begin{cases} 0, & \text{si } i \neq j, \\ 1, & \text{si } i = j. \end{cases}$$

Por tanto, cuando $i \neq j$,

$$H_{n,j}(x_i) = 0 \text{ y } \hat{H}_{n,j}(x_i) = 0,$$

mientras que

$$H_{n,i}(x_i) = [1 - 2(x_i - x_i)L'_{n,i}(x_i)] \cdot 1 = 1$$

y

$$\hat{H}_{n,i}(x_i) = (x_i - x_i) \cdot 1^2 = 0.$$

En consecuencia,

$$H_{2n+1}(x_i) = \sum_{\substack{j=0 \\ j \neq i}}^{n} f(x_j) \cdot 0 + f(x_i) \cdot 1 + \sum_{j=0}^{n} f'(x_j) \cdot 0 = f(x_i),$$

así que H_{2n+1} concuerda con f en $x_0, x_1, \ldots, x_n$.

Si queremos demostrar la concordancia de H'_{2n+1} con f' en los nodos, primero observamos que $L_{n,j}(x)$ es un factor de $H'_{n,j}(x)$, de modo que $H'_{n,j}(x_i) = 0$ cuando $i \neq j$. Además, si $i = j$ y $L_{n,i}(x_i) = 1$, tenemos,

$$\begin{aligned} H'_{n,i}(x_i) &= -2L'_{n,i}(x_i) \cdot L^2_{n,i}(x_i) + [1 - 2(x_i - x_i)L'_{n,i}(x_i)]2L_{n,i}(x_i)L'_{n,i}(x_i) \\ &= -2L'_{n,i}(x_i) + 2L'_{n,i}(x_i) = 0. \end{aligned}$$

Por tanto $H'_{n,j}(x_i) = 0$ para todas las i y las j.

Finalmente,

$$\begin{aligned} \hat{H}'_{n,j}(x_i) &= L^2_{n,j}(x_i) + (x_i - x_j)2L_{n,j}(x_i)L'_{n,j}(x_i) \\ &= L_{n,j}(x_i)[L_{n,j}(x_i) + 2(x_i - x_j)L'_{n,j}(x_i)], \end{aligned}$$

así que $\hat{H}'_{n,j}(x_i) = 0$ si $i \neq j$ y $\hat{H}'_{n,i}(x_i) = 1$. Al combinar estos hechos tenemos

$$H'_{2n+1}(x_i) = \sum_{j=0}^{n} f(x_j) \cdot 0 + \sum_{\substack{j=0 \\ j \neq i}}^{n} f'(x_j) \cdot 0 + f'(x_i) \cdot 1 = f'(x_i).$$

Por tanto, H_{2n+1} concuerda con f y H'_{2n+1} con f' en $x_0, x_1, \ldots, x_n$.

En el ejercicio 8 se consideran la unicidad de este polinomio y la fórmula de error.

■ ■ ■

EJEMPLO 1 Utilice el polinomio de Hermite que concuerda con los datos de la tabla 3.12 para obtener una aproximación de $f(1.5)$.

Tabla 3.12

k	x_k	$f(x_k)$	$f'(x_k)$
0	1.3	0.6200860	−0.5220232
1	1.6	0.4554022	−0.5698959
2	1.9	0.2818186	−0.5811571

Calcule primero los polinomios de Lagrange y sus derivadas:

$$L_{2,0}(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}, \qquad L'_{2,0}(x) = \frac{100}{9}x - \frac{175}{9};$$

$$L_{2,1}(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}, \qquad L'_{2,1}(x) = \frac{-200}{9}x + \frac{320}{9};$$

y

$$L_{2,2} = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}, \qquad L'_{2,2}(x) = \frac{100}{9}x - \frac{145}{9}.$$

Así, los polinomios $H_{2,j}(x)$ y $\hat{H}_{2,j}(x)$ son

$$H_{2,0}(x) = [1 - 2(x-1.3)(-5)]\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2$$

$$= (10x - 12)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2,$$

$$H_{2,1}(x) = 1 \cdot \left(\frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2,$$

$$H_{2,2}(x) = 10(2-x)\left(\frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}\right)^2,$$

$$\hat{H}_{2,0}(x) = (x-1.3)\left(\frac{50}{9}x^2 - \frac{175}{9}x + \frac{152}{9}\right)^2,$$

$$\hat{H}_{2,1}(x) = (x-1.6)\left(\frac{-100}{9}x^2 + \frac{320}{9}x - \frac{247}{9}\right)^2,$$

y

$$\hat{H}_{2,2}(x) = (x-1.9)\left(\frac{50}{9}x^2 - \frac{145}{9}x + \frac{104}{9}\right)^2.$$

Finalmente,

$$H_5(x) = 0.6200860H_{2,0}(x) + 0.4554022H_{2,1}(x) + 0.2818186H_{2,2}(x)$$

$$- 0.5220232\hat{H}_{2,0}(x) - 0.5698959\hat{H}_{2,1}(x) - 0.5811571\hat{H}_{2,2}(x)$$

y

$$H_5(1.5) = 0.6200860\left(\frac{4}{27}\right) + 0.4554022\left(\frac{64}{81}\right) + 0.2818186\left(\frac{5}{81}\right)$$
$$- 0.5220232\left(\frac{4}{405}\right) - 0.5698959\left(\frac{-32}{405}\right) - 0.5811571\left(\frac{-2}{405}\right)$$
$$= 0.5118277,$$

resultado cuya exactitud corresponde a los lugares decimales indicados. ■

Aunque el teorema 3.9 proporciona una descripción completa de los polinomios de Hermite, en el ejemplo 1 comprobamos lo siguiente: la necesidad de determinar y evaluar los polinomios de Lagrange y sus derivadas hace tedioso el procedimiento, aun para valores pequeños de n. Otro método para generar las aproximaciones de Hermite tiene como base la fórmula de diferencias divididas interpolantes de Newton (3.10) para el polinomio de Lagrange en $x_0, x_1, \ldots, x_n$,

$$P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0) \cdots (x - x_{k-1}),$$

y la conexión entre la n-ésima diferencia dividida y la derivada de grado n de f como se describió en el teorema 3.6 de la sección 3.2.

Supongamos que los números distintos $x_0, x_1, \ldots, x_n$, están dados junto con los valores de f y de f' en esos números. Defina una sucesión nueva $z_0, z_1, \ldots, z_{2n+1}$ por medio de

$$z_{2i} = z_{2i+1} = x_i, \qquad \text{para cada } i = 0, 1, \ldots, n.$$

y construya la tabla de diferencias divididas en la forma de la tabla 3.7 que utiliza $z_0, z_1, \ldots, z_{2n+1}$.

Puesto que $z_{2i} = z_{2i+1} = x_i$ para cada i, no podemos definir $f[z_{2i}, z_{2i+1}]$ a partir de la fórmula de diferencias divididas. Si suponemos, basándonos en el teorema 3.6, que la sustitución razonable en este caso es $f[z_{2i}, z_{2i+1}] = f'(z_{2i}) = f'(x_i)$, podremos utilizar las entradas o datos

$$f'(x_0), f'(x_1), \ldots, f'(x_n)$$

en vez de las primeras diferencias divididas no definidas

$$f[z_0, z_1], f[z_2, z_3], \ldots, f[z_{2n}, z_{2n+1}].$$

Las diferencias divididas restantes se obtienen en la forma habitual y las diferencias divididas apropiadas se emplean en la fórmula de diferencias divididas interpolantes de Newton. La tabla 3.13 contiene los datos que se emplean en las columnas de las tres primeras diferencias divididas cuando se determina el polinomio de Hermite $H_5(x)$ para x_0, x_1 y x_2. Los datos restantes se generan tal como se hizo en la tabla 3.7. El polinomio de Hermite está dado por

$$H_{2n+1}(x) = f[z_0] + \sum_{k=1}^{2n+1} f[z_0, \ldots, z_k](x - z_0)(x - z_1) \cdots (x - z_{k-1}).$$

Una demostración de este hecho viene en [Po, p. 56].

Tabla 3.13

z	$f(z)$	Primeras diferencias divididas	Segundas diferencias divididas
$z_0 = x_0$	$f[z_0] = f(x_0)$		
		$f[z_0, z_1] = f'(x_0)$	
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2] = \dfrac{f[z_1, z_2] - f[z_0, z_1]}{z_2 - z_0}$
		$f[z_1, z_2] = \dfrac{f[z_2] - f[z_1]}{z_2 - z_1}$	
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3] = \dfrac{f[z_2, z_3] - f[z_1, z_2]}{z_3 - z_1}$
		$f[z_2, z_3] = f'(x_1)$	
$z_3 = x_1$	$f[z_3] = f(x_1)$		$f[z_2, z_3, z_4] = \dfrac{f[z_3, z_4] - f[z_2, z_3]}{z_4 - z_2}$
		$f[z_3, z_4] = \dfrac{f[z_4] - f[z_3]}{z_4 - z_3}$	
$z_4 = x_2$	$f[z_4] = f(x_2)$		$f[z_3, z_4, z_5] = \dfrac{f[z_4, z_5] - f[z_3, z_4]}{z_5 - z_3}$
		$f[z_4, z_5] = f'(x_2)$	
$z_5 = x_2$	$f[z_5] = f(x_2)$		

EJEMPLO 2 Los valores de la tabla 3.14 usan los datos del ejemplo 1. Los valores subrayados son los datos conocidos; los restantes se generan mediante la fórmula de las diferencias divididas ordinarias (3.9):

$$\begin{aligned} H_5(1.5) &= 0.6200860 + (1.5 - 1.3)(-0.5220232) + (1.5 - 1.3)^2(-0.0897427) \\ &\quad + (1.5 - 1.3)^2(1.5 - 1.6)(0.0663657) + (1.5 - 1.3)^2(1.5 - 1.6)^2(0.0026663) \\ &\quad + (1.5 - 1.3)^2(1.5 - 1.6)^2(1.5 - 1.9)(-0.0027738) \\ &= 0.5118277. \end{aligned}$$

■

Tabla 3.14

1.3	0.6200860					
		−0.5220232				
1.3	0.6200860		−0.0897427			
		−0.5489460		0.0663657		
1.6	0.4554022		−0.0698330		0.0026663	
		−0.5698959		0.0679655		−0.0027738
1.6	0.4554022		−0.0290537		0.0010020	
		−0.5786120		0.0685667		
1.9	0.2818186		−0.0084837			
		−0.5811571				
1.9	0.2818186					

La técnica aplicada en el algoritmo 3.3 puede ampliarse y usarse para determinar otros polinomios osculantes. Una explicación concisa de los procedimientos la encuentra en [Po, pp. 53-57].

ALGORITMO 3.3

Interpolación de Hermite

Para obtener los coeficientes del polinomio interpolante de Hermite $H(x)$ en los $(n + 1)$ números distintos $x_0, \ldots, x_n$ para la función f:

ENTRADA los números $x_0, x_1, \ldots, x_n$; valores $f(x_0), \ldots, f(x_n)$ y $f'(x_0), \ldots, f'(x_n)$.

SALIDA los números $Q_{0,0}, Q_{1,1}, \ldots, Q_{2n+1,2n+1}$ donde

$$H(x) = Q_{0,0} + Q_{1,1}(x - x_0) + Q_{2,2}(x - x_0)^2 + Q_{3,3}(x - x_0)^2(x - x_1) + Q_{4,4}(x - x_0)^2(x - x_1)^2 + \cdots + Q_{2n+1,2n+1}(x - x_0)^2(x - x_1)^2 \cdots (x - x_{n-1})^2(x - x_n).$$

Paso 1 Para $i = 0, 1, \ldots, n$ haga pasos 2 y 3.

Paso 2 Sea $z_{2i} = x_i$;
$z_{2i+1} = x_i$;
$Q_{2i,0} = f(x_i)$;
$Q_{2i+1,0} = f(x_i)$;
$Q_{2i+1,1} = f'(x_i)$.

Paso 3 Si $i \neq 0$ entonces tome

$$Q_{2i,1} = \frac{Q_{2i,0} - Q_{2i-1,0}}{z_{2i} - z_{2i-1}}.$$

Paso 4 Para $i = 2, 3, \ldots, 2n + 1$

para $j = 2, 3, \ldots, i$ tomar $Q_{i,j} = \dfrac{Q_{i,j-1} - Q_{i-1,j-1}}{z_i - z_{i-j}}$.

Paso 5 ENTRADA $(Q_{0,0}, Q_{1,1}, \ldots, Q_{2n+1,2n+1})$;
PARAR. ■

CONJUNTO DE EJERCICIOS 3.3

1. Use el teorema 3.9 o el algoritmo 3.3 para construir un polinomio de aproximación para los siguientes datos.

a.

x	$f(x)$	$f'(x)$
8.3	17.56492	3.116256
8.6	18.50515	3.151762

b.

x	$f(x)$	$f'(x)$
0.8	0.22363362	2.1691753
1.0	0.65809197	2.0466965

c.

x	$f(x)$	$f'(x)$
−0.5	−0.0247500	0.7510000
−0.25	0.3349375	2.1890000
0	1.1010000	4.0020000

d.

x	$f(x)$	$f'(x)$
0.1	−0.62049958	3.58502082
0.2	−0.28398668	3.14033271
0.3	0.00660095	2.66668043
0.4	0.24842440	2.16529366

2. Los datos del ejercicio 1 se generaron por medio de las siguientes funciones. Use los polinomios construidos en el ejercicio 1 para el valor dado de x y con ellos aproxime $f(x)$ y calcule el error real.

a. $f(x) = x \ln x$; aproximar $f(8.4)$.
b. $f(x) = \text{sen}(e^x - 2)$; aproximar $f(0.9)$.
c. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$; aproximar $f(-\frac{1}{3})$.
d. $f(x) = x \cos x - 2x^2 + 3x - 1$; aproximar $f(0.25)$.

3. a. Use los siguientes valores y la aritmética de redondeo a cinco dígitos para construir un polinomio interpolante de Hermite que le permita aproximar sen 0.34.

x	sen x	D_x sen x = cos x
0.30	0.29552	0.95534
0.32	0.31457	0.94924
0.35	0.34290	0.93937

b. Determine una cota de error para la aproximación de la parte (a) y compárela con el error real.

c. Agregue sen 0.33 = 0.32404 y cos 0.33 = 0.94604 a los datos y vuelva a efectuar los cálculos.

4. Sea $f(x) = 3xe^x - e^{2x}$.

a. Aproxime $f(1.03)$ por medio del polinomio interpolante de Hermite de grado máximo tres utilizando $x_0 = 1$ y $x_1 = 1.05$. Compare el error real con la cota del error.

b. Repita (a) con el polinomio interpolante de Hermite de grado máximo cinco, utilizando $x_0 = 1$, $x_1 = 1.05$ y $x_2 = 1.07$.

5. Use la fórmula de error y Maple para encontrar una cota de los errores en las aproximaciones de $f(x)$ en las partes (a) y (c) del ejercicio 2.

6. La tabla siguiente contiene datos referentes a la función que se describe mediante $f(x) = e^{0.1x^2}$. Aproxime $f(1.25)$ por medio de $H_5(1.25)$ y $H_3(1.25)$, donde H_5 usa los nodos $x_0 = 1$, $x_1 = 2$ y $x_2 = 3$ y H_3 emplean los nodos $\bar{x}_0 = 1$ y $\bar{x}_1 = 1.5$. Calcule las cotas de error en estas aproximaciones.

x	$f(x) = e^{0.1x^2}$	$f'(x) = 0.2xe^{0.1x^2}$
$x_0 = \bar{x}_0 = 1$	1.105170918	0.2210341836
$\bar{x}_1 = 1.5$	1.252322716	0.3756968148
$x_1 = 2$	1.491824698	0.5967298792
$x_2 = 3$	2.459603111	1.475761867

7. Un automóvil realiza un recorrido por una carretera recta y se cronometra su recorrido en varios puntos. Los datos recabados de las observaciones se incluyen en la tabla adjunta, donde el tiempo se indica en segundos, la distancia en pies y la velocidad en pies por segundo.

Tiempo	0	3	5	8	13
Distancia	0	225	383	623	993
Velocidad	75	77	80	74	72

a. Use el polinomio de Hermite para predecir la posición del automóvil y su velocidad cuando $t = 10$ s.

b. Use la derivada del polinomio de Hermite para determinar si el automóvil rebasa el límite de velocidad de 55 mi/h en la carretera. De ser así, ¿cuál es la primera vez que la excede?

c. ¿Cuál es la velocidad máxima predecible del automóvil?

8. a. Demuestre que $H_{2n+1}(x)$ es el polinomio único de menor grado que concuerda con f y con f' en $x_0, \ldots, x_n$. [*Sugerencia*: suponga que $P(x)$ es otro polinomio de este tipo y considere $D = H_{2n+1} - P$ y D' en $x_0, x_1, \ldots, x_n$.]

b. Deduzca el término de error en el teorema 3.9 [*Sugerencia:* aplique el mismo método que en la deducción del error de Lagrange, teorema 3.3, que define

$$g(t) = f(t) - H_{2n+1}(t) - \frac{(t-x_0)^2 \cdots (t-x_n)^2}{(x-x_0)^2 \cdots (x-x_n)^2}\,[f(x) - H_{2n+1}(x)]$$

y demuestre que $g'(t)$ tiene $(2n + 2)$ ceros distintos en $[a, b]$.]

9. Sean $z_0 = x_0$, $z_1 = x_0$, $z_2 = x_1$ y $z_3 = x_1$. Construya la tabla anexa de diferencias divididas.

$z_0 = x_0$	$f[z_0] = f(x_0)$			
		$f[z_0, z_1] = f'(x_0)$		
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2]$	
		$f[z_1, z_2]$		$f[z_0, z_1, z_2, z_3]$
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3]$	
		$f[z_2, z_3] = f'(x_1)$		
$z_3 = x_1$	$f[z_3] = f(x_1)$			

Demuestre que el polinomio cúbico de Hermite $H_3(x)$ también puede reescribirse como $f[z_0] + f[z_0, z_1](x - x_0) + f[z_0, z_1, z_2](x - x_0)^2 + f[z_0, z_1, z_2, z_3](x - x_0)^2(x - x_1)$.

3.4 Interpolación de trazadores cúbicos*

En secciones anteriores de este capítulo estudiamos la aproximación de una función arbitraria por medio de un polinomio en un intervalo cerrado. Sin embargo, la naturaleza oscilatoria de los polinomios de alto grado y la propiedad de que una fluctuación en una parte pequeña de un intervalo puede ocasionar importantes fluctuaciones en todo el rango limita su utilización. Al final de esta sección veremos un buen ejemplo de ello. (Véase la Fig. 3.12.)

Un procedimiento alterno consiste en dividir el intervalo en una serie de subintervalos, y en cada subintervalo construir un polinomio (generalmente) diferente de aproximación. A esta forma de aproximar por medio de funciones se le conoce como **aproximación polinómica fragmentaria.**

La aproximación polinómica fragmentaria es la interpolación lineal fragmentaria que consiste en unir una serie de puntos de datos

$$\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$$

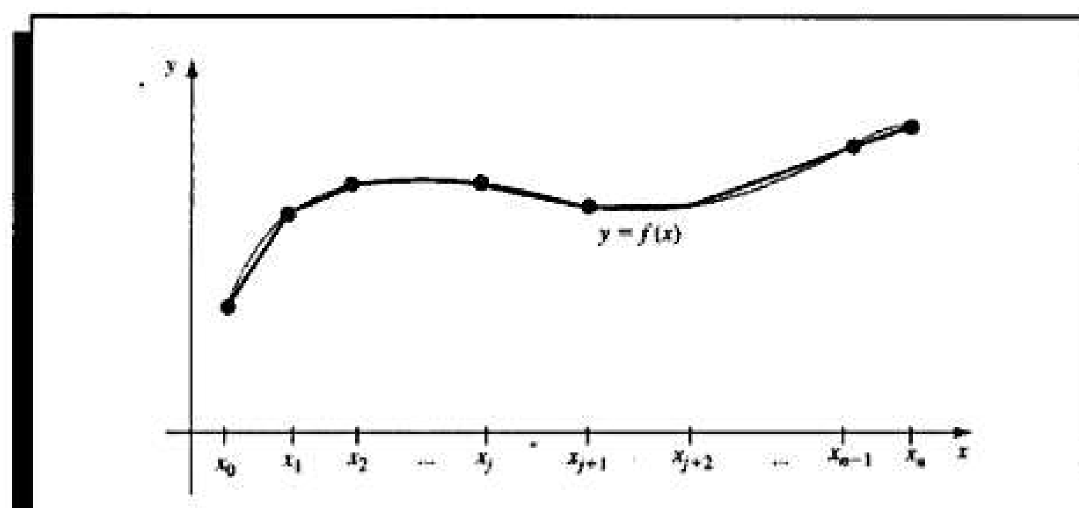
mediante una serie de segmentos de rectas, como los que aparecen en la figura 3.7.

La aproximación por funciones lineales muestra una desventaja: no se tiene la seguridad de que haya diferenciabilidad en los extremos de los subintervalos, lo cual dentro de un contexto geométrico significa que la función interpolante no es "suave" en dichos puntos. A menudo las condiciones físicas indican claramente que se requiere esa condición y que la función aproximante debe ser continuamente diferenciable.

Otro procedimiento consiste en emplear un polinomio fragmentario del tipo Hermite. Por ejemplo, si los valores de la función f y de f' se conocen en los puntos $x_0 < x_1 < \ldots < x_n$, podemos emplear un polinomio de Hermite de grado tres en cada uno de los subintervalos $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$ para obtener una función continuamente diferenciable en el intervalo $[x_0, x_n]$. Si queremos determinar el polinomio cúbico Hermite apropiado en

* Las demostraciones de los teoremas de esta sección se basan en los resultados del capítulo 6.

Figura 3.7



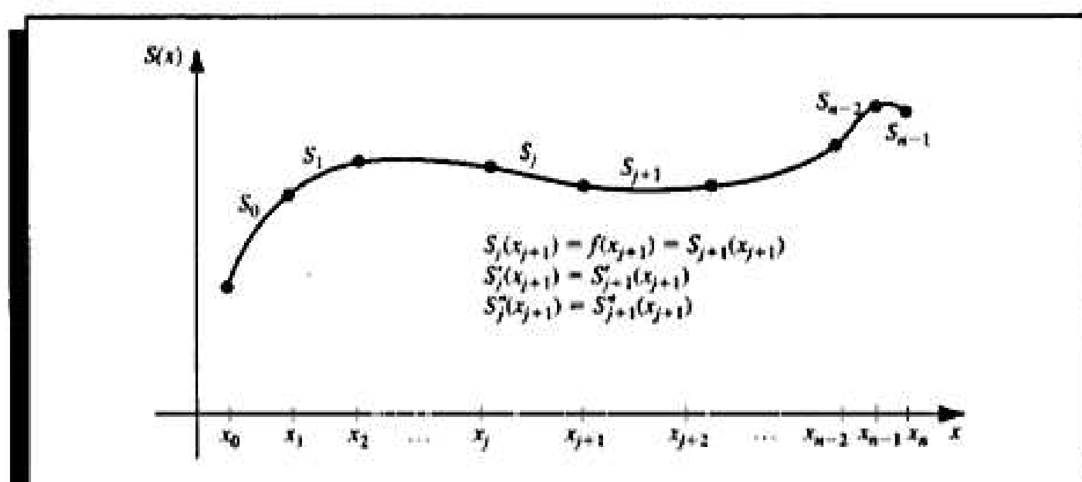
determinado intervalo, basta calcular $H_3(x)$ para ese intervalo. Puesto que los polinomios interpolantes de Lagrange necesarios para calcular H_3 son de primer grado, podemos hacer el cálculo sin gran dificultad. Sin embargo, para utilizar los polinomios fragmentarios de Hermite en la interpolación general, necesitamos conocer la derivada de la función que va a ser aproximada, lo cual muchas veces no es posible.

En lo que resta de esta sección estudiaremos la aproximación por medio de polinomios fragmentarios que no requieren información sobre la derivada, salvo, quizá, en los extremos del intervalo donde se aproxima la función.

El tipo más simple de función de polinomio fragmentario diferenciable en un intervalo entre $[x_0, x_n]$ es la función obtenida al ajustar un polinomio cuadrático entre cada par consecutivo de nodos. Esto se hace construyendo una cuadrática en $[x_0, x_1]$ que concuerde con la función en x_0 y en x_1; y así sucesivamente. Un polinomio cuadrático general tiene tres constantes arbitrarias —el término constante, el coeficiente de x y el coeficiente de x^2— y únicamente se requieren dos condiciones para ajustar los datos en los extremos de cada intervalo, por ello, existe flexibilidad que permite seleccionar la cuadrática de modo que la interpolante tenga una derivada continua en $[x_0, x_n]$. El problema de este procedimiento se presenta cuando hay que especificar las condiciones referentes a la derivada de la interpolante en los extremos x_0 y x_n. No hay constantes suficientes para cerciorarse de que se satisfagan las condiciones (véase el ejercicio 22).

La aproximación polinómica fragmentaria más común utiliza polinomios entre cada par consecutivo de nodos y recibe el nombre de **interpolación de trazadores cúbicos**. Un polinomio cúbico general contiene cuatro constantes; así pues, el procedimiento del trazador cúbico ofrece suficiente flexibilidad para garantizar que el interpolante no sólo sea continuamente diferenciable en el intervalo, sino que además tenga una segunda derivada continua en el intervalo. Sin embargo, en la construcción del trazador cúbico no se supone que las derivadas del interpolante concuerdan con las de la función, ni siquiera en los nodos. (Véase la Fig. 3.8.)

Figura 3.8



Definición 3.10 Dada una función f definida en $[a, b]$ y un conjunto de nodos $a = x_0 < x_1 < \cdots < x_n = b$ un **interpolante de trazador cúbico** S para f es una función que cumple con las condiciones siguientes:

a. $S(x)$ es un polinomio cúbico, denotado $S_j(x)$, en el subintervalo $[x_j, x_{j+1}]$ para cada $j = 0, 1, \ldots, n-1$;

b. $S(x_j) = f(x_j)$ para cada $j = 0, 1, \ldots, n$;

c. $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para cada $j = 0, 1, \ldots, n-2$;

d. $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ para cada $j = 0, 1, \ldots, n-2$;

e. $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ para cada $j = 0, 1, \ldots, n-2$;

f. Una de las siguientes condiciones de frontera se satisface:

(i) $S''(x_0) = S''(x_n) = 0$ **(frontera libre o natural)**;

(ii) $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$ **(frontera sujeta)**.

Aunque los trazadores cúbicos se definen con otras condiciones de frontera, las condiciones dadas en (f) son suficientes en este caso. Cuando se presentan las condiciones de frontera libre, el trazador recibe el nombre de **trazador natural** y su gráfica se aproxima a la forma que adoptaría una varilla larga y flexible si la hiciéramos pasar por los puntos $\{(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$.

En términos generales, en las condiciones de frontera sujeta se logran aproximaciones más exactas, ya que abarcan más información acerca de la función. Pero para que se cumpla este tipo de condición de frontera, se requiere tener los valores de la derivada en los extremos o bien una aproximación precisa de ellos.

Si queremos construir el interpolante del trazador cúbico de determinada función f, aplicamos las condiciones de la definición a los polinomios cúbicos:

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3,$$

para cada $j = 0, 1, \ldots, n - 1$.

Está claro que

$$S_j(x_j) = a_j = f(x_j),$$

y si se aplica la condición (c),

$$a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3,$$

para cada $j = 0, 1, \ldots, n - 2$.

Puesto que los términos $x_{j+1} - x_j$ se utilizarán varias veces en este desarrollo, conviene introducir la notación más simple

$$h_j = x_{j+1} - x_j,$$

para cada $j = 0, 1, \ldots, n - 1$. Si también definimos $a_n = f(x_n)$, entonces la ecuación

$$a_{j+1} = a_j + b_jh_j + c_jh_j^2 + d_jh_j^3 \tag{3.15}$$

será válida para cada $j = 0, 1, \ldots, n - 1$.

De manera análoga, defina $b_n = S'(x_n)$ y observe que

$$S_j'(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$$

significa que $S_j'(x_j) = b_j$ para cada $j = 0, 1, \ldots, n - 1$. Al aplicar la condición (d) obtenemos

$$b_{j+1} = b_j + 2c_jh_j + 3d_jh_j^2, \tag{3.16}$$

para cada $j = 0, 1, \ldots, n - 1$.

Al definir $c_n = S''(x_n)/2$ y aplicar la condición (e), se obtiene otra relación entre los coeficientes de S_j. En este caso, para cada $j = 0, 1, \ldots, n - 1$.

$$c_{j+1} = c_j + 3d_jh_j. \tag{3.17}$$

Al despejar d_j en la ecuación (3.17) y sustituir este valor en las ecuaciones (3.15) y (3.16), para cada $j = 0, 1, \ldots, n - 1$ se obtienen las ecuaciones

$$a_{j+1} = a_j + b_jh_j + \frac{h_j^2}{3}(2c_j + c_{j+1}) \tag{3.18}$$

y

$$b_{j+1} = b_j + h_j(c_j + c_{j+1}) \tag{3.19}$$

La relación final que incluye los coeficientes se obtiene resolviendo la ecuación correspondiente en la forma de la ecuación (3.18), primero para b_j,

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \tag{3.20}$$

y entonces, con una reducción del índice, para b_{j-1}:

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j).$$

Cuando sustituimos estos valores en la ecuación obtenida de la ecuación (3.19), con el índice reducido en 1, obtenemos el sistema de ecuaciones lineales

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}), \tag{3.21}$$

para cada $j = 1, 2, \ldots, n - 1$. Este sistema contiene sólo $\{c_j\}_{j=0}^{n}$ como incógnitas, ya que los valores de $\{h_j\}_{j=0}^{n-1}$ y de $\{a_j\}_{j=0}^{n}$ están dados por el espaciado de los nodos $\{x_j\}_{j=0}^{n}$ y los valores de f en éstos.

Nótese que una vez que se conocen los valores de $\{c_j\}_{j=0}^{n}$, encontrar el resto de las constantes $\{b_j\}_{j=0}^{n-1}$ partiendo de la ecuación (3.20) y $\{d_j\}_{j=0}^{n-1}$ de la ecuación (3.17) para construir los polinomios cúbicos $\{S_j(x)\}_{j=0}^{n-1}$ es fácil.

El interrogante principal que se plantea en relación con esta construcción es si se pueden determinar los valores de $\{c_j\}_{j=0}^{n}$ por medio del sistema de ecuaciones dado en (3.21) y, de ser así, si estos valores son únicos. El siguiente teorema indica que esto es posible cuando se establece una de las dos condiciones de frontera de la parte (f) de la definición. Las demostraciones de estos teoremas requieren los conceptos de álgebra lineal que se explican en el capítulo 6.

Teorema 3.11 Si definimos f en $a = x_0 < x_1 < \cdots < x_n = b$ entonces f tendrá un interpolante único de trazador natural en los nodos $x_0, x_1, \ldots, x_n$; es decir, un interpolante de trazador que cumple con las condiciones de frontera $S''(a) = 0$ y $S''(b) = 0$. ■

Demostración En este caso, las condiciones de frontera significan que $c_n = S''(x_n)/2 = 0$ y que

$$0 = S''(x_0) = 2c_0 + 6d_0(x_0 - x_0);$$

así que $c_0 = 0$.

Las dos ecuaciones $c_0 = 0$ y $c_n = 0$ junto con las ecuaciones de (3.21) producen un sistema lineal descrito por la ecuación vectorial $A\mathbf{x} = \mathbf{b}$, donde A es la matriz $(n + 1) \times (n + 1)$

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ddots & & & \vdots \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix},$$

y donde $\mathbf{b}$ y $\mathbf{x}$ son los vectores

$$\mathbf{b} = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \quad \text{y} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

La matriz A es estrictamente dominante en sentido diagonal, de manera que satisface las hipótesis del teorema 6.19 de la sección 6.6. Por tanto, el sistema lineal tiene una solución única para $c_0, c_1, \ldots, c_n$. ■ ■ ■

La solución del problema de los trazadores cúbicos con las condiciones de frontera $S''(x_0) = S''(x_n) = 0$ puede obtenerse por medio del algoritmo 3.4.

ALGORITMO 3.4

Trazador cúbico natural

Para construir el interpolante de trazador cúbico S de la función f, que se define en los números $x_0 < x_1 < \cdots < x_n$ y que satisface $S''(x_0) = S''(x_n) = 0$:

ENTRADA n; $x_0, x_1, \ldots, x_n$; $a_0 = f(x_0), a_1 = f(x_1), \ldots, a_n = f(x_n)$.

SALIDA a_j, b_j, c_j, d_j, para $j = 0, 1, \ldots, n-1$.
(*Nota*: $S(x) = S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *para* $x_j \leq x \leq x_{j+1}$.)

Paso 1 Para $i = 0, 1, \ldots, n-1$ tome $h_i = x_{i+1} - x_i$.

Paso 2 Para $i = 1, 2, \ldots, n-1$ tome

$$\alpha_i = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}).$$

Paso 3 Tome $l_0 = 1$; (*Los pasos 3, 4, 5 y parte del paso 6 resuelven un sistema lineal tridiagonal utilizando el método descrito en el algoritmo 6.7.*)
$\mu_0 = 0$;
$z_0 = 0$.

Paso 4 Para $i = 1, 2, \ldots, n-1$
tome $l_i = 2(x_{i+1} - x_{i-1}) - h_{i-1}\mu_{i-1}$;
$\mu_i = h_i / l_i$;
$z_i = (\alpha_i - h_{i-1}z_{i-1})/l_i$.

Paso 5 Tome $l_n = 1$;
$z_n = 0$;
$c_n = 0$.

Paso 6 Para $j = n-1, n-2, \ldots, 0$
tome $c_j = z_j - \mu_j c_{j+1}$;
$b_j = (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3$;
$d_j = (c_{j+1} - c_j)/(3h_j)$.

Paso 7 SALIDA (a_j, b_j, c_j, d_j, para $j = 0, 1, \ldots, n-1$);
PARAR. ■

En el caso de las condiciones de frontera sujeta se obtiene un resultado similar al del teorema 3.11.

Teorema 3.12 Si f está definida en $a = x_0 < x_1 < \cdots < x_n = b$, y es diferenciable en a y b, entonces f tiene un único trazador sujeto que interpola los nodos $x_0, x_1, \ldots, x_n$, es decir, un interpolante de trazador que cumple las condiciones de frontera $S'(a) = f'(a)$ y $S'(b) = f'(b)$. ■

Demostración Puesto que $f(a) = S'(a) = S'(x_0) = b_0$, podemos ver que la ecuación (3.20) con $j = 0$ implica que

$$f'(a) = \frac{1}{h_0}(a_1 - a_0) - \frac{h_0}{3}(2c_0 + c_1).$$

En consecuenica,

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(a).$$

De manera semejante,

$$f'(b) = b_n = b_{n-1} + h_{n-1}(c_{n-1} + c_n),$$

de modo que la ecuación (3.20) con $j = n - 1$ implica que

$$f'(b) = \frac{a_n - a_{n-1}}{h_{n-1}} - \frac{h_{n-1}}{3}(2c_{n-1} + c_n) + h_{n-1}(c_{n-1} + c_n)$$

$$= \frac{a_n - a_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{3}(c_{n-1} + 2c_n),$$

y que

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}).$$

Las ecuaciones (3.21) junto con las ecuaciones

$$2h_0c_0 + h_0c_1 = \frac{3}{h_0}(a_1 - a_0) - 3f'(a)$$

y

$$h_{n-1}c_{n-1} + 2h_{n-1}c_n = 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1})$$

determinan el sistema lineal $A\mathbf{x} = \mathbf{b}$, donde

$$A = \begin{bmatrix} 2h_0 & h_0 & 0 & \cdots & \cdots & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & \ddots & & & \vdots \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & h_{n-2} & 2(h_{n-2} + h_{n-1}) & & h_{n-1} \\ 0 & \cdots & \cdots & 0 & & h_{n-1} & 2h_{n-1} \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} \frac{3}{h_0}(a_1 - a_0) - 3f'(a) \\ \frac{3}{h_1}(a_2 - a_1) = \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}) \end{bmatrix} \quad \text{y} \quad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

La matriz A es estrictamente dominante en sentido diagonal y, por tanto, cumple con las condiciones del teorema 6.19. En consecuencia, el sistema lineal tiene una solución única para $c_0, c_1, \ldots, c_n$. ■ ■ ■

La solución del problema de los trazadores cúbicos con condiciones de frontera $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$ se puede obtener usando el algoritmo 3.5.

ALGORITMO 3.5

Trazador cúbico sujeto

Para construir el interpolante de trazado cúbico S para la función f que se define en los números $x_0 < x_1 < \cdots < x_n$, y que satisface $S'(x_0) = f'(x_0)$ y $S'(x_n) = f'(x_n)$:

ENTRADA $n; x_0, x_1, \ldots, x_n; a_0 = f(x_0), a_1 = f(x_1), \ldots, a_n = f(x_n)$; $FPO = f'(x_0)$; $FPN = f'(x_n)$.

SALIDA a_j, b_j, c_j, d_j, para $j = 0, 1, \ldots, n - 1$.
(*Nota:* $S(x) = S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *para* $x_j \leq x \leq x_{j+1}$.)

Paso 1 Para $i = 0, 1, \ldots, n - 1$ tome $h_i = x_{i+1} - x_i$.

Paso 2 Tome $\alpha_0 = 3(a_1 - a_0)/h_0 - 3\,FPO$;
$\alpha_n = 3\,FPN - 3(a_n - a_{n-1})/h_{n-1}$

Paso 3 Para $i = 1, 2, \ldots, n - 1$

$$\textit{tome } \alpha_i = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}).$$

Paso 4 Tome $l_0 = 2h_0$; (*Los Pasos 4, 5, 6 y parte del paso 7 resuelven un sistema lineal tridiagonal utilizando un método descrito en el algoritmo 6.7.*)
$\mu_0 = 0.5$;
$z_0 = \alpha_0 / l_0$.

Paso 5 Para $i = 1, 2, \ldots, n-1$
tome $l_i = 2(x_{i+1} - x_{i-1}) - h_{i-1}\mu_{i-1}$;
$\mu_i = h_i / l_i$;
$z_i = (\alpha_i - h_{i-1}z_{i-1}) / l_i$.

Paso 6 Tome $l_n = h_{n-1}(2 - \mu_{n-1})$;
$z_n = (\alpha_n - h_{n-1}z_{n-1}) / l_n$;
$c_n = z_n$.

Paso 7 Para $j = n-1, n-2, \ldots, 0$
tome $c_j = z_j - \mu_j c_{j+1}$;
$b_j = (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3$;
$d_j = (c_{j+1} - c_j)/(3h_j)$.

Paso 8 SALIDA (a_j, b_j, c_j, d_j para $j = 0, 1, \ldots, n-1$);
PARAR. ■

EJEMPLO 1 La figura 3.9 muestra a un joven pato en pleno vuelo. Para aproximar el perfil de la parte superior del pato, seleccionamos algunos puntos a lo largo de la curva por donde queremos que pase la curva de aproximación. La tabla 3.15 incluye las coordenadas de 21 puntos de datos relativos al sistema de coordenadas sobrepuestas que aparece en la figura 3.10. Obsérvese que se utilizan más puntos cuando la curva cambia rápidamente que cuando lo hace con más lentitud.

Figura 3.9

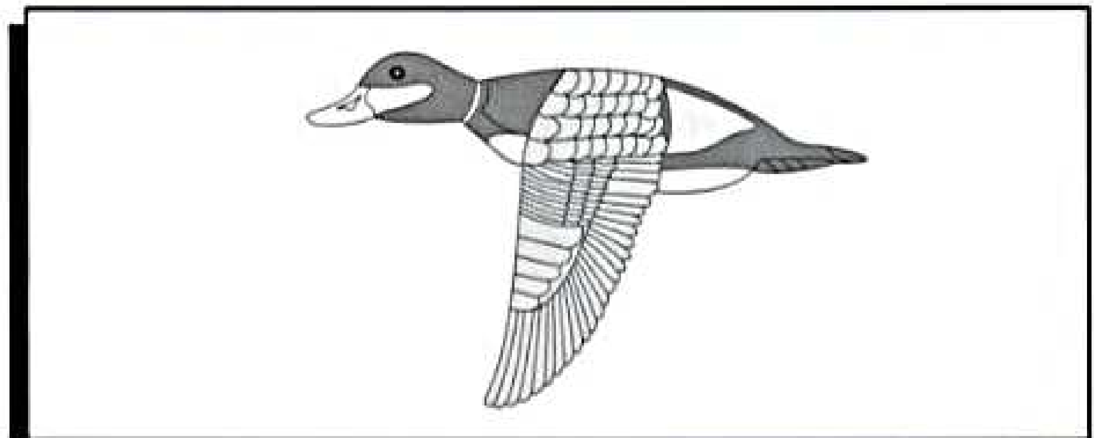


Tabla 3.15

x	0.9	1.3	1.9	2.1	2.6	3.0	3.9	4.4	4.7	5.0	6.0	7.0	8.0	9.2	10.5	11.3	11.6	12.0	12.6	13.0	13.3
$f(x)$	1.3	1.5	1.85	2.1	2.6	2.7	2.4	2.15	2.05	2.1	2.25	2.3	2.25	1.95	1.4	0.9	0.7	0.6	0.5	0.4	0.25

Al utilizar el algoritmo 3.4 para generar el trazador cúbico libre con estos datos, se obtienen los coeficientes que aparecen en la tabla 3.16. Esta curva de trazador es casi idéntica al perfil, como se observa en la figura 3.11.

Figura 3.10

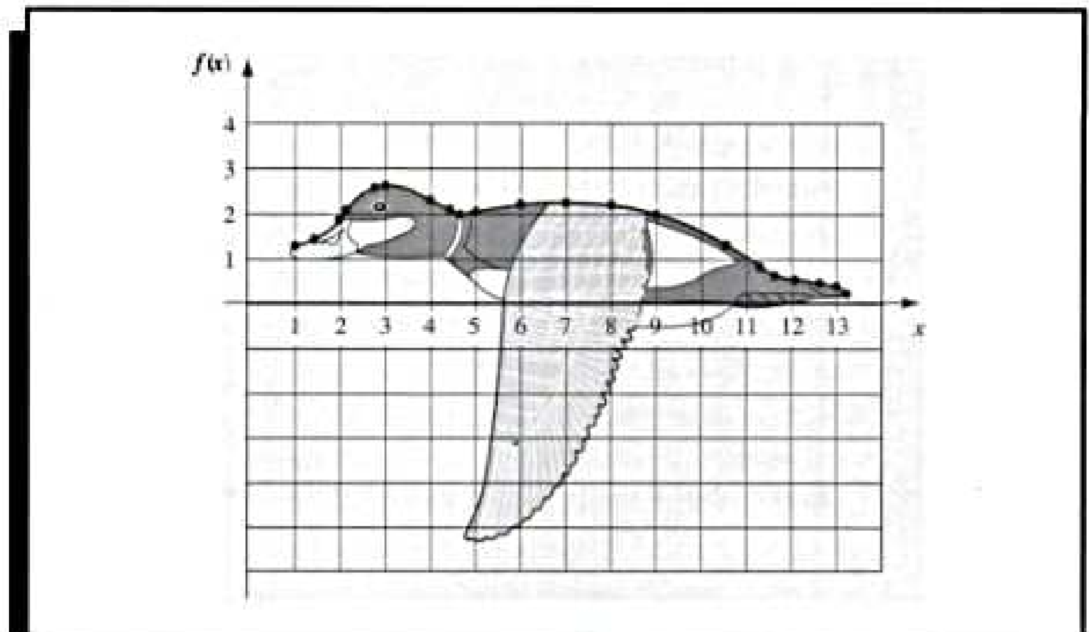
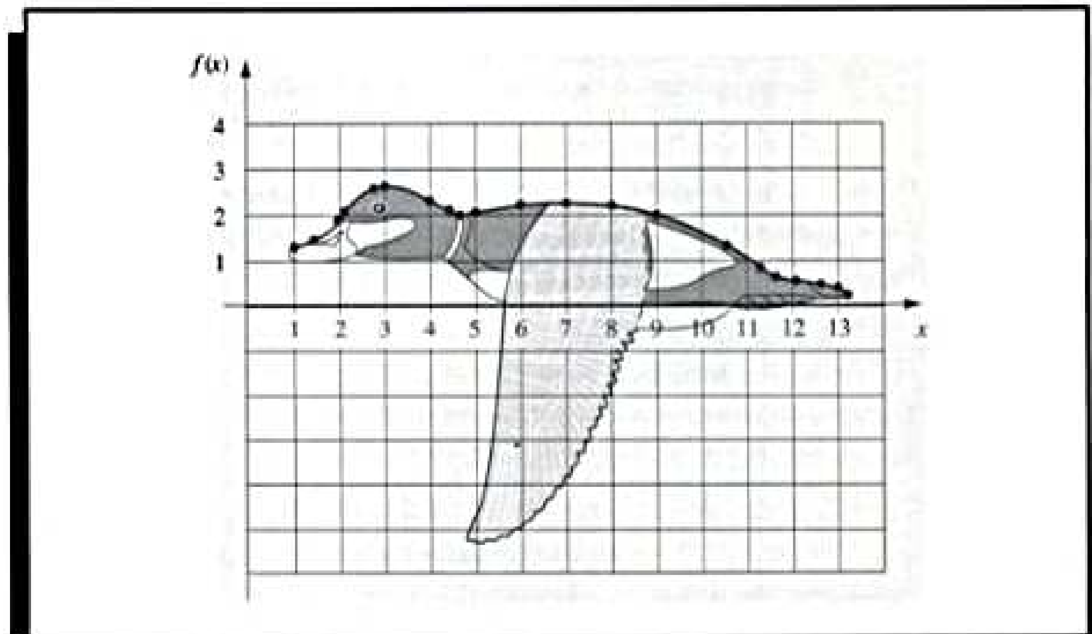


Tabla 3.16

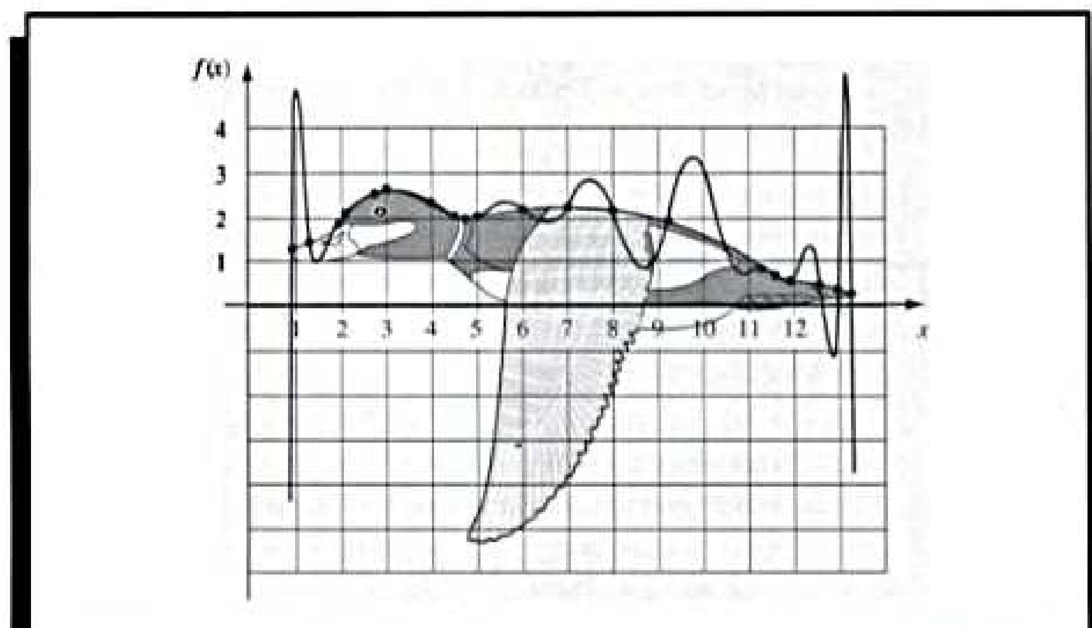
j	x_j	a_j	b_j	c_j	d_j
0	0.9	1.3	5.40	0.00	−0.25
1	1.3	1.5	0.42	−0.30	0.95
2	1.9	1.85	1.09	1.41	−2.96
3	2.1	2.1	1.29	−0.37	−0.45
4	2.6	2.6	0.59	−1.04	0.45
5	3.0	2.7	−0.02	−0.50	0.17
6	3.9	2.4	−0.50	−0.03	0.08
7	4.4	2.15	−0.48	0.08	1.31
8	4.7	2.05	−0.07	1.27	−1.58
9	5.0	2.1	0.26	−0.16	0.04
10	6.0	2.25	0.08	−0.03	0.00
11	7.0	2.3	0.01	−0.04	−0.02
12	8.0	2.25	−0.14	−0.11	0.02
13	9.2	1.95	−0.34	−0.05	−0.01
14	10.5	1.4	−0.53	−0.10	−0.02
15	11.3	0.9	−0.73	−0.15	1.21
16	11.6	0.7	−0.49	0.94	−0.84
17	12.0	0.6	−0.14	−0.06	0.04
18	12.6	0.5	−0.18	0.00	−0.45
19	13.0	0.4	−0.39	−0.54	0.60
20	13.3	0.25			

Figura 3.11



A fin de facilitar la comparación, la figura 3.12 ofrece un ejemplo gráfico de la curva que se genera con un polinomio interpolante de Lagrange que corresponde a los datos de la tabla 3.15, generándose así una extraña ilustración del lomo del pato en vuelo o en otra posición. En este caso, el polinomio interpolante es de grado 20 y oscila demasiado.

Figura 3.12



Si queremos utilizar un trazador sujeto para aproximar esta curva, necesitaremos aproximaciones de la derivada para los extremos. Y aun cuando tuviéramos dichas aproximaciones, podríamos esperar poco o nada de mejora a causa de la estrecha concordancia entre el trazador cúbico libre y la curva del perfil de la parte superior. ■

Sería mucho más difícil construir un trazador cúbico para aproximar el perfil de la parte inferior del pato, ya que la curva de esta parte no puede expresarse en función de x y en algunos puntos la curva no parece ser suave. Estos problemas se pueden resolver por medio de trazadores individuales que representan varias partes de la curva, pero en la siguiente sección estudiaremos un procedimiento más eficaz para las curvas de este tipo.

Por lo general es preferible contar con las condiciones de frontera sujeta cuando se aproximan funciones por medio de trazadores cúbicos; de ahí la necesidad de estimar la derivada de la función en los extremos del intervalo. En el caso en que los nodos estén igualmente espaciados cerca de ambos extremos, podemos obtener las aproximaciones mediante la ecuación (4.7) o cualquier fórmula apropiada de las que vienen en las secciones 4.1 y 4.2. En el caso de nodos espaciados de forma desigual, el problema resulta mucho más difícil.

A manera de conclusión para esta sección, incluiremos una fórmula para la cota de error del trazador cúbico con condiciones de frontera sujeta. La demostración de este resultado viene en [Schul, pp. 57-58].

Teorema 3.13 Sea $f \in C^4[a, b]$ con $\max_{a \le x \le b} |f^{(4)}(x)| = M$. Si S es el interpolante único de trazador cúbico sujeto de f respecto a los nodos $a = x_0 < x_1 < \cdots < x_n = b$, entonces

$$\max_{a \le x \le b} |f(x) - S(x)| \le \frac{5M}{384} \max_{0 \le j \le n-1} (x_{j+1} - x_j)^4. \quad ■$$

Un resultado para la cota de error del cuarto orden también se obtiene en el caso de las condiciones de frontera libre, pero es más difícil de expresar. (Véase [BD, pp. 827-835].)

Cerca de los extremos del intervalo $[x_0, x_n]$, las condiciones de frontera libre casi siempre dan resultados menos exactos a menos que la función f casualmente satisfaga $f''(x_0) = f''(x_n) = 0$. Una alternativa de la condición de frontera libre que no requiere el conocimiento de la derivada de f es la condición *no un nodo* (véase [Deb, pp. 55-56]). Esta condición exige que $S'''(x)$ sea continuo en x_1 y x_{n-1}.

CONJUNTO DE EJERCICIOS 3.4

1. Determine el trazador cúbico libre S que interpola los datos $f(0) = 0, f(1) = 1$ y $f(2) = 2$.
2. Determine el trazador cúbico sujeto s que interpola los datos $f(0) = 0, f(1) = 1, f(2) = 2$ y que satisface $s'(0) = s'(2) = 1$.
3. Construya el trazador cúbico libre de los siguientes datos.

a.

x	$f(x)$
8.3	17.56492
8.6	18.50515

b.

x	$f(x)$
0.8	0.22363362
1.0	0.65809197

c.

x	$f(x)$
-0.5	-0.0247500
-0.25	0.3349375
0	1.1010000

d.

x	$f(x)$
0.1	-0.62049958
0.2	-0.28398668
0.3	0.00660095
0.4	0.24842440

4. Los datos del ejercicio 3 se generaron usando las siguientes funciones. Utilice los trazadores cúbicos construidos en el ejercicio 3 para el valor dado de x, a fin de aproximar $f(x)$ y $f'(x)$. También calcule el error real.

a. $f(x) = x \ln x$; aproxime $f(8.4)$ y $f'(8.4)$.

b. $f(x) = \operatorname{sen}(e^x - 2)$; aproxime $f(0.9)$ y $f'(0.9)$.

c. $f(x) = x^3 + 4.001x^2 + 4.002x + 1.101$; aproxime $f(-\frac{1}{3})$ y $f'(-\frac{1}{3})$.

d. $f(x) = x \cos x - 2x^2 + 3x - 1$; aproxime $f(0.25)$ y $f'(0.25)$.

5. Construya el trazador cúbico sujeto aplicando los datos del ejercicio 3 y el hecho de que

a. $f'(8.3) = 1.116256$ y $f'(8.6) = 1.151762$

b. $f'(0.8) = 2.1691753$ y $f'(1.0) = 2.0466965$

c. $f'(-0.5) = 0.7510000$ y $f'(0) = 4.0020000$

d. $f'(0.1) = 3.58502082$ y $f'(0.4) = 2.16529366$

6. Repita el ejercicio 4 por medio de los trazadores cúbicos construidos en el ejercicio 5.

7. Un trazador cúbico natural S en $[0, 2]$ está definido por

$$S(x) = \begin{cases} S_0(x) = 1 + 2x - x^3, & \text{si } 0 \le x < 1, \\ S_1(x) = 2 + b(x-1) + c(x-1)^2 + d(x-1)^3, & \text{si } 1 \le x \le 2. \end{cases}$$

Obtenga b, c y d.

8. Un trazador cúbico sujeto s de la función f está definido en $[1, 3]$ por

$$s(x) = \begin{cases} s_0(x) = 3(x-1) + 2(x-1)^2 - (x-1)^3, & \text{si } 1 \le x < 2, \\ s_1(x) = a + b(x-2) + c(x-2)^2 + d(x-2)^3, & \text{si } 2 \le x \le 3. \end{cases}$$

Dadas $f'(1) = f'(3)$, encuentre a, b, c y d.

9. Un trazador cúbico natural S está definido por

$$S(x) = \begin{cases} S_0(x) = 1 + B(x-1) - D(x-1)^3, & \text{si } 1 \le x < 2, \\ S_1(x) = 1 + b(x-2) - \frac{3}{4}(x-2)^2 + d(x-2)^3, & \text{si } 2 \le x \le 3. \end{cases}$$

Si S interpola los datos $(1, 1)$, $(2, 1)$ y $(3, 0)$, obtenga B, D, b y d.

10. Un trazador cúbico sujeto s de la función f está definido por

$$s(x) = \begin{cases} s_0(x) = 1 + Bx + 2x^2 - 2x^3, & \text{si } 0 \le x < 1, \\ s_1(x) = 1 + b(x-1) - 4(x-1)^2 + 7(x-1)^3, & \text{si } 1 \le x \le 2. \end{cases}$$

Obtenga $f'(0)$ y $f'(2)$.

11. Construya un trazado cúbico libre para aproximar $f(x) = \cos \pi x$ utilizando para ello los valores que se dan en $f(x)$ en $x = 0, 0.25, 0.5, 0.75$ y 1.0. Integre el trazador en $[0, 1]$ y compare el resultado con $\int_0^1 \cos \pi x dx = 0$. Aproxime $f'(0.5)$ y $f''(0.5)$ por medio de las derivadas del trazador. Compare esas aproximaciones con los valores reales.

12. Construya un trazador cúbico libre para aproximar $f(x) = e^{-x}$ por medio de los valores dados por $f(x)$ en $x = 0, 0.25, 0.75$ y 1.0. Integre el trazador en $[0, 1]$ y compare el resultado con

$\int_0^1 e^{-x}\,dx = 1 - 1/e$. Mediante las derivadas del trazador aproxime $f'(0.5)$ y $f''(0.5)$. Compare las aproximaciones con los valores reales.

13. Repita el ejercicio 11, construyendo esta vez el trazado cúbico sujeto con $f'(0) = f'(1) = 0$.

14. Repita el ejercicio 12, construyendo esta vez el trazador cúbico sujeto con $f'(0) = -1, f'(1) = -e^{-1}$.

15. Suponga que $f(x)$ es un polinomio de grado 3. Demuestre que $f(x)$ es su propio trazador cúbico sujeto, pero que no puede ser su propio trazador cúbico libre.

16. Suponga que los datos $\{(x_i, f(x_i))\}_{i=1}^n$ se encuentran sobre una recta. ¿Qué podemos decir sobre los trazadores cúbicos libres y sujetos de la función f? [*Sugerencia*: aproveche los resultados de los ejercicios 1 y 2.]

17. Dada la partición $x_0 = 0$, $x_1 = 0.05$ y $x_2 = 0.1$ de $[0, 0.1]$, obtenga la función interpolante lineal fragmentaria F para $f(x) = e^{2x}$. Aproxime $\int_0^{0.1} e^{2x}\,dx$ con $\int_0^{0.1} F(x)\,dx$, y compare el resultado con el valor real.

18. Sea $f \in C^2[a, b]$, y sean $a = x_0 < x_1 < \cdots < x_n = b$ los nodos dados. Deduzca una estimación de error semejante a la del teorema 3.13 para la función interpolante lineal fragmentaria F. Por medio de esta estimación, derive las cotas de error correspondientes al ejercicio 17.

19. Extienda los algoritmos 3.4 y 3.5 para incluir como salida las derivadas primera y segunda del trazador en los nodos.

20. Extienda los algoritmos 3.4 y 3.5 para incluir como salida la integral del trazador en el intervalo $[x_0, x_n]$.

21. Dada la partición $x_0 = 0$, $x_1 = 0.05$, $x_2 = 0.1$ de $[0, 0.1]$ y $f(x) = e^{2x}$:

   a. Obtenga el trazador cúbico s con las condiciones de frontera sujeta que interpola f.

   b. Obtenga una aproximación de $\int_0^{0.1} e^{2x}\,dx$ evaluando $\int_0^{0.1} s(x)\,dx$.

   c. Aplique el teorema 3.13 para estimar $\max_{0\le x\le 0.1} |f(x) - s(x)|$ y

$$\left|\int_0^{0.1} f(x)\,dx - \int_0^{0.1} s(x)\,dx\right|.$$

   d. Determine el trazador cúbico S con condiciones de frontera libre y compare después $S(0.02)$, $s(0.02)$ y $e^{0.04} = 1.04081077$.

22. Sea f la función definida en $[a, b]$ y sean $a = x_0 < x_1 < x_2 = b$ los nodos dados. Una función interpolante del trazador cuadrático S está compuesto por el polinomio cuadrático

$$S_0(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 \quad \text{en } [x_0, x_1],$$

y el polinomio cuadrático

$$S_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 \quad \text{en } [x_1, x_2],$$

tales que

(i) $S(x_0) = f(x_0)$, $S(x_1) = f(x_1)$ y $S(x_2) = f(x_2)$,

(ii) $S \in C^1[x_0, x_2]$.

Demuestre que las condiciones (i) y (ii) dan origen a cinco ecuaciones en las seis incógnitas a_0, b_0, c_0, a_1, b_1 y c_1. El problema consiste en decidir qué condición adicional establecer para hacer que la solución sea única. ¿Se obtiene una solución significativa con la condición $S \in C^2[x_0, x_2]$

23. Determine un trazado cuadrático s que interpole los datos $f(0) = 0$, $f(1) = 1$, $f(2) = 2$ y que satisfaga $s'(0) = 2$.

24. **a.** En la introducción del capítulo se incluyó una tabla que contiene la población de Estados Unidos durante el periodo comprendido entre 1940 y 1990. Por medio de la interpolación de trazador cúbico libre, aproxime la población en los años 1930, 1965 y 2010.

b. En 1930, la población fue aproximadamente de 123 203 000 habitantes. ¿Qué exactitud, en su opinión, tienen las estadísticas correspondientes a los años 1965 y 2010?

25. Un automóvil va por una carretera recta y su velocidad se cronometra en varios puntos. Los datos tomados de las observaciones aparecen en la tabla adjunta, donde el tiempo se anota en segundos, la distancia en pies y la velocidad en pies por segundo.

Tiempo	0	3	5	8	13
Distancia	0	225	383	623	993
Velocidad	75	77	80	74	72

a. Use un trazador cúbico sujeto para predecir la posición del automóvil y su velocidad cuando $t = 10$ s.

b. Use la derivada del trazador para determinar si el automóvil rebasa el límite de velocidad de 55 mi/h; de ser así, ¿en qué momento el automóvil lo excede?

c. ¿Cuál es la velocidad máxima predecible del automóvil?

26. El caballo llamado Thunder Gulch ganó el Derby de Kentucky de 1995, con un tiempo de 2:01 $\frac{1}{5}$ (2 min y 1 $\frac{1}{5}$ s) en la carrera de 1 $\frac{1}{4}$ millas. Los tiempos en los postes que marcan el cuarto de milla, la mitad de la milla y la milla fueron, respectivamente, 22 $\frac{2}{5}$, 45 $\frac{4}{5}$ y 1:35 $\frac{3}{5}$.

a. Use los valores anteriores junto con el tiempo de arranque y construya un trazador cúbico libre para la carrera de Thunder Gulch.

b. Use el trazador para predecir el tiempo en el poste de tres cuartos de milla y compare el resultado con el tiempo real de 1:10 $\frac{1}{5}$.

c. Use el trazador para aproximar la velocidad inicial de Thunder Gulch y la velocidad en la meta.

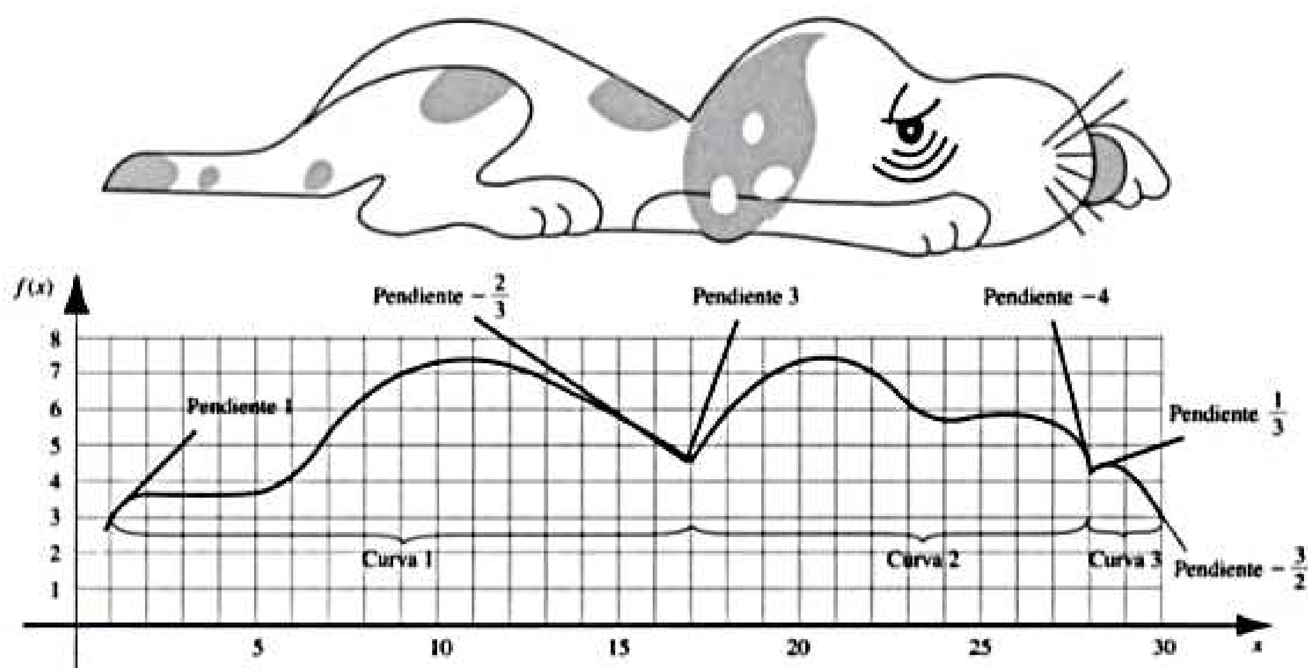
27. Se sospecha que las concentraciones altas de tanina en las hojas maduras de roble inhiben el crecimiento de larvas de la polilla de invierno (*Operophtera bromata L., Geometridae*) que tanto dañan a estos árboles en algunos años. En la lista anexa se incluye el peso promedio de dos muestras de larvas en los primeros 28 días después del nacimiento. La primera muestra se crió en hojas de roble joven y la segunda en hojas maduras del mismo árbol.

a. Por medio del trazador cúbico libre aproxime la curva de peso promedio de cada muestra.

b. Obtenga un peso promedio máximo aproximado de cada muestra, determinando para ello el máximo del trazador.

Día	0	6	10	13	17	20	28
Peso promedio de la muestra 1 (mg)	6.67	17.33	42.67	37.33	30.10	29.31	28.74
Peso promedio de la muestra 2 (mg)	6.67	16.11	18.89	15.00	10.56	9.44	8.89

28. Mediante los interpolantes de trazadores cúbicos sujetos se va a aproximar la parte superior del perro que se muestra en la siguiente página. Se traza la curva en una cuadrícula a partir de la cual se construyó la tabla. Utilice el algoritmo 3.5 para construir los tres trazadores cúbicos sujetos.



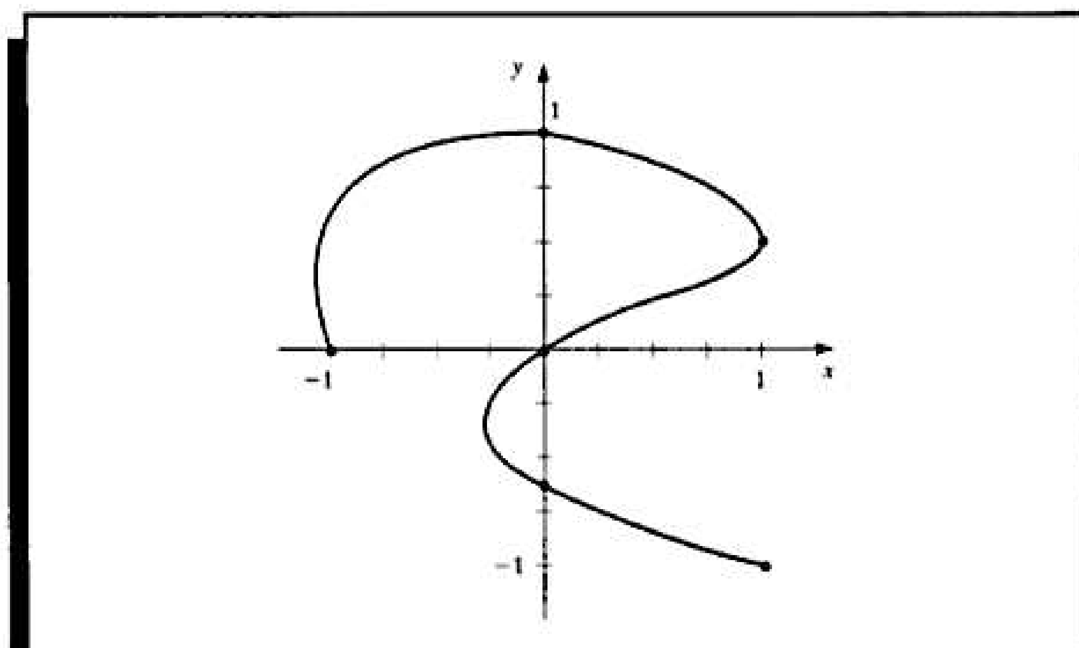
Curva 1				Curva 2				Curva 3			
i	x_i	$f(x_i)$	$f'(x_i)$	i	x_i	$f(x_i)$	$f'(x_i)$	i	x_i	$f(x_i)$	$f'(x_i)$
0	1	3.0	1.0	0	17	4.5	3.0	0	27.7	4.1	0.33
1	2	3.7		1	20	7.0		1	28	4.3	
2	5	3.9		2	23	6.1		2	29	4.1	
3	6	4.2		3	24	5.6		3	30	3.0	−1.5
4	7	5.7		4	25	5.8					
5	8	6.6		5	27	5.2					
6	10	7.1		6	27.7	4.1	−4.0				
7	13	6.7									
8	17	4.5	−0.67								

29. Repita el ejercicio 28 construyendo tres trazadores naturales con el algoritmo 3.4.

3.5 Curvas paramétricas

Ninguna de las técnicas que hemos desarrollado sirve para generar curvas como la que aparece en la figura 3.13, ya que esta curva no puede expresarse en función de una variable coordenada a partir de la otra. En esta sección explicaremos cómo representar curvas generales aplicando un parámetro para expresar las variables de las coordenadas x y y. Este procedimiento se puede aplicar también para representar las curvas y superficies generales en el espacio.

Figura 3.13



Un método paramétrico sencillo con el que se determina un polinomio o un polinomio fragmentario para conectar los puntos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, en el orden dado consiste en usar un parámetro t en un intervalo $[t_0, t_n]$ con $t_0 < t_1 < \cdots < t_n$, y construir las funciones de aproximación mediante

$$x_i = x(t_i) \quad \text{y} \quad y_i = y(t_i) \quad \text{para cada } i = 0, 1, \ldots, n.$$

El siguiente ejemplo demuestra la técnica en el caso en que las dos funciones aproximantes sean polinomios interpolantes de Lagrange:

EJEMPLO 1 Construya un par de polinomios de Lagrange para aproximar la curva de la figura 3.13, empleando los puntos de datos que aparecen en la curva.

La selección del parámetro admite flexibilidad, y escogeremos los puntos $\{t_i\}_{i=0}^{n}$ igualmente espaciados en [0, 1]. En este caso, contamos con los datos de la tabla 3.17.

Tabla 3.17

i	0	1	2	3	4
t_i	0	0.25	0.5	0.75	1
x_i	−1	0	1	0	1
y_i	0	1	0.5	0	−1

Esto genera los polinomios interpolantes

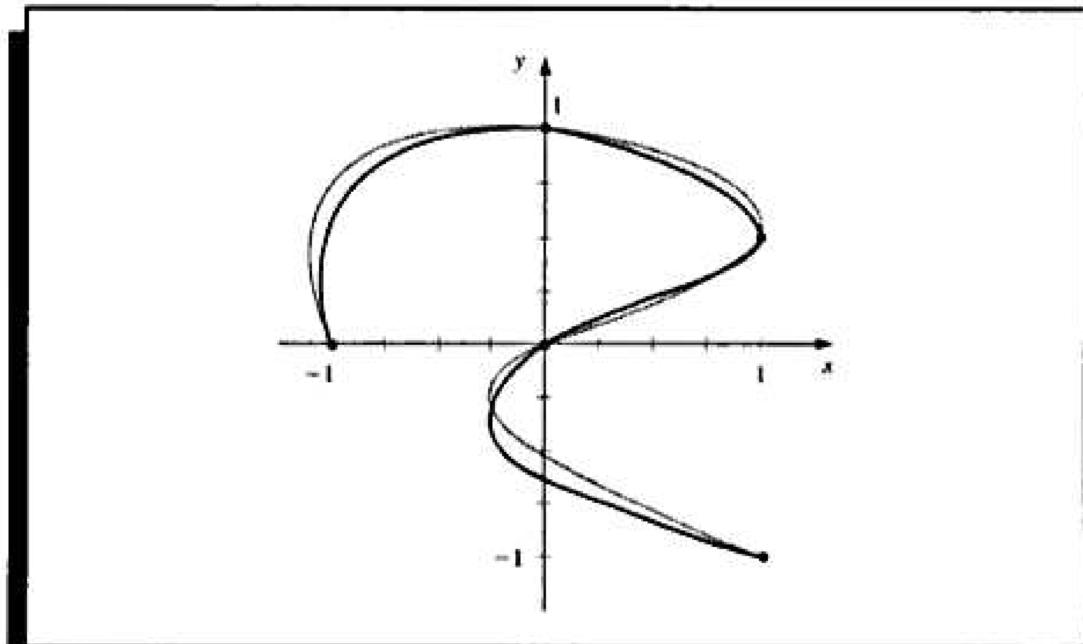
$$x(t) = \left(\left(\left(64t - \frac{352}{3}\right)t + 60\right)t - \frac{14}{3}\right)t - 1$$

y

$$y(t) = \left(\left(\left(-\frac{64}{3}t + 48\right)t - \frac{116}{3}\right)t + 11\right)t.$$

Al graficar este sistema paramétrico se obtiene la gráfica de la figura 3.14. Aunque esta curva pasa por los puntos requeridos y presenta la misma forma básica, es una aproximación burda a la curva original. Una aproximación más exacta requeriría más nodos, con el consecuente aumento de cálculos. ■

Figura 3.14



Las curvas de Hermite y del trazador pueden generarse de modo semejante, pero una vez más debemos efectuar muchos cálculos.

Las aplicaciones de las gráficas por computadora requieren la rápida generación de curvas suaves que puedan ser fácil y rápidamente modificables. Además, por razones estéticas y de cálculo, el cambio de una parte de estas curvas debe tener poco o nulo efecto en otras partes. Así debemos prescindir de los polinomios y trazadores interpolantes, pues al cambiar una parte de las curvas influimos en la curva entera.

En general, la selección de la curva que se utilizará en la gráfica por computadora es una forma del polinomio cúbico fragmentario de Hermite. Todas las partes de un polinomio de este tipo se determinan íntegramente al especificar sus extremos y las derivadas en éstos. En consecuencia, es posible modificar una parte de la curva sin alterar la mayor parte de ella; sólo se deben modificar las partes adyacentes para garantizar un aspecto suave en los extremos. Los cálculos se pueden efectuar rápidamente y se puede modificar una sección de la curva a la vez.

El problema que implica la interpolación de Hermite es la necesidad de especificar las derivadas en los extremos de cada sección de la curva. Supongamos que la curva tiene $n + 1$ puntos de datos $(x(t_0), y(t_0)), \ldots, (x(t_n), y(t_n))$ y que se desea parametrizar el cúbico para incluir características complejas. Entonces se debe especificar $x'(t_i)$ y $y'(t_i)$ para cada $i = 0, 1, \ldots, n$. Ello no resulta tan difícil como parece a primera vista, puesto que ca-

da parte puede generarse de modo independiente, a condición de que nos aseguremos de que las derivadas en los extremos de cada porción corresponden a la parte adyacente. Así pues, en esencia se puede simplificar el procedimiento y convertirlo simplemente en determinar, en cada sección de la curva, un par de los polinomios de Hermite en el parámetro t, donde $t_0 = 0$ y $t_1 = 1$, dados los datos de los extremos $(x(0), y(0))$, $(x(1), y(1))$ y las derivadas dy/dx (en $t = 0$) y dy/dx (en $t = 1$).

Obsérvese, no obstante, que estamos especificando sólo seis condiciones y que cada polinomio cúbico tiene cuatro parámetros, lo cual nos da un total de ocho. Esta situación permite una gran flexibilidad al seleccionar el par de polinomios cúbicos de Hermite para cumplir con estas condiciones. La flexibilidad se debe a que la forma natural de determinar $x(t)$ y $y(t)$ requiere especificar $x'(0)$, $x'(1)$, $y'(0)$ y $y'(1)$. La curva explícita de Hermite en x y en y requiere especificar sólo los cocientes

$$\frac{dy}{dx}(t = 0) = \frac{y'(0)}{x'(0)} \quad \text{y} \quad \frac{dy}{dx}(t = 1) = \frac{y'(1)}{x'(1)}.$$

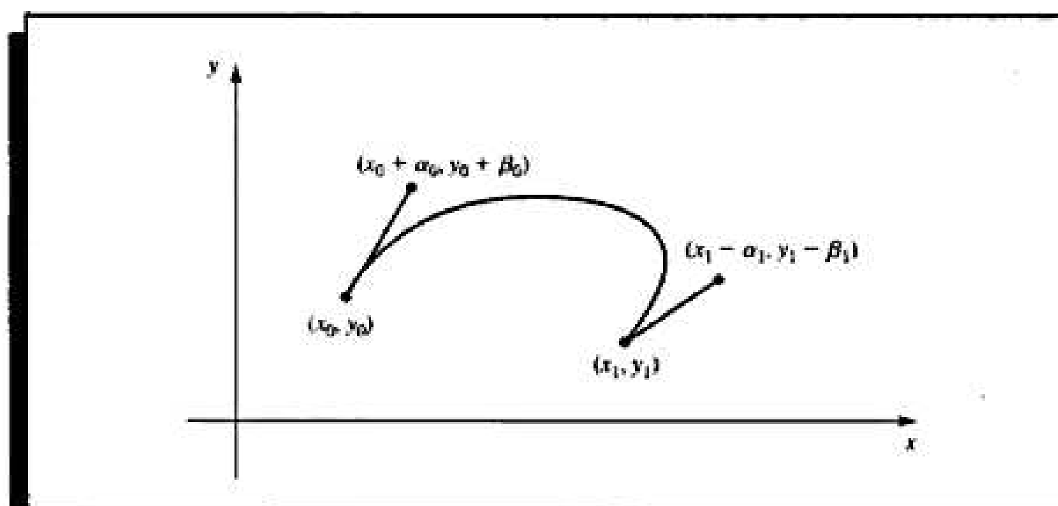
Al multiplicar $x'(0)$ y $y'(0)$ por un factor escalar común, la tangente de la curva en $(x(0), y(0))$ permanece sin cambio, pero la forma de la curva sí varía. Cuando mayor sea el factor escalar, más cerca estará la curva de aproximar la tangente cerca de $(x(0), y(0))$. Una situación parecida se presenta en el otro extremo $(x(1), y(1))$.

Si queremos simplificar aún más el proceso en las gráficas interactivas por computadora, hay que especificar gráficamente la derivada en un extremo y describir un segundo punto, denominado *punto guía*, en la tangente deseada. Cuanto más lejos esté ese punto del nodo, mayor será el factor escalar y la curva aproximará más la tangente cerca del nodo.

En la figura 3.15 los nodos se encuentran en (x_0, y_0) y en (x_1, y_1), el punto guía de (x_0, y_0) es $(x_0 + \alpha_0, y_0 + \beta_0)$, y el de (x_1, y_1) es $(x_1 + \alpha_1, y_1 - \beta_1)$. El polinomio cúbico de Hermite $x(t)$ en $[0, 1]$ debe satisfacer

$$x(0) = x_0, \qquad x(1) = x_1, \qquad x'(0) = \alpha_0 \qquad \text{y} \qquad x'(1) = \alpha_1.$$

Figura 3.15



Se puede verificar fácilmente que el polinomio cúbico único que cumple con estas condiciones es

$$x(t) = [2(x_0 - x_1) + (\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - (\alpha_1 + 2\alpha_0)]t^2 + \alpha_0 t + x_0. \qquad (3.22)$$

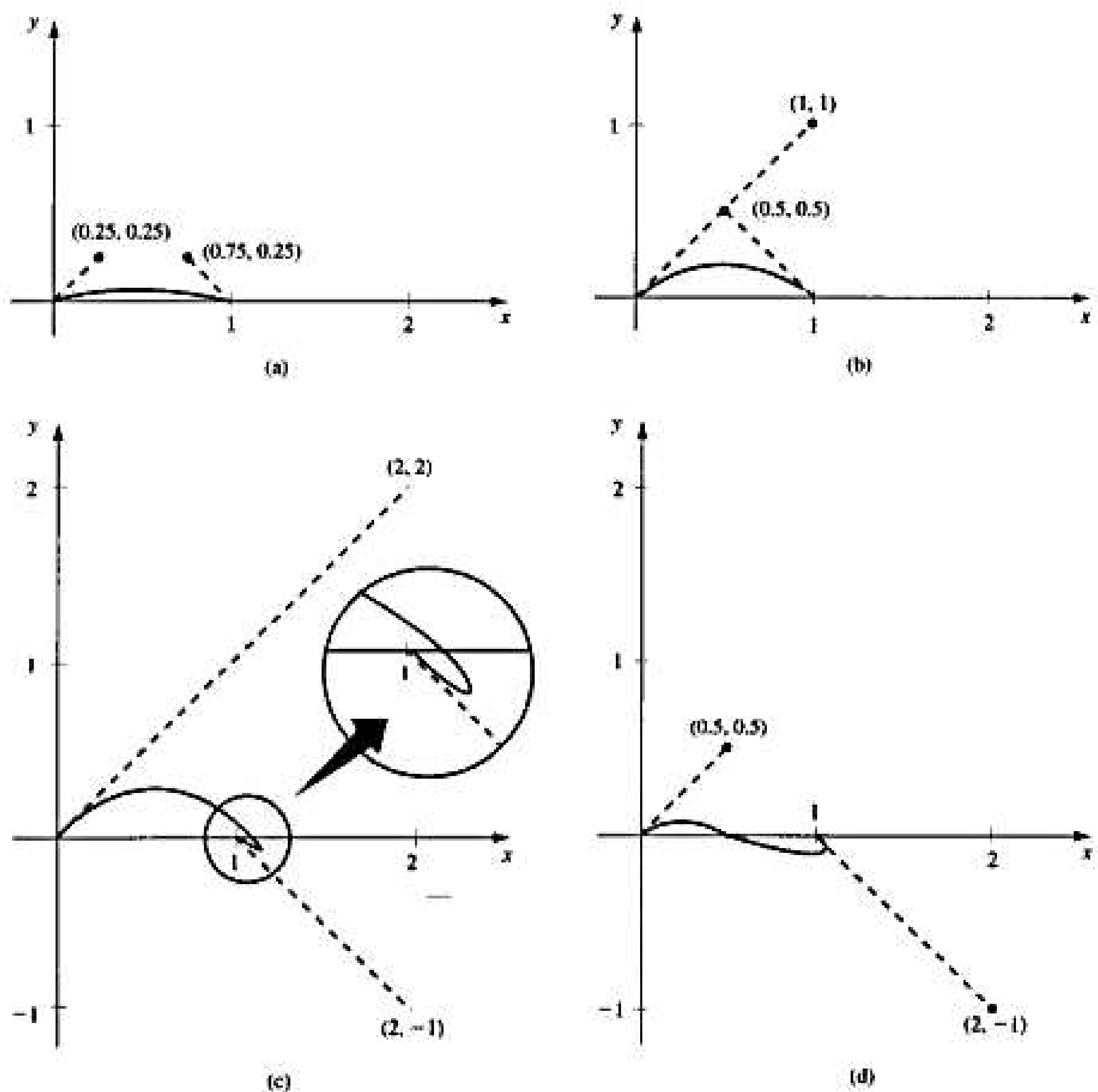
De modo análogo, el polinomio cúbico único que satisface

$$y(0) = y_0, \quad y(1) = y_1, \quad y'(0) = \beta_0 \quad \text{y} \quad y'(1) = \beta_1$$

es

$$y(t) = [2(y_0 - y_1) + (\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - (\beta_1 + 2\beta_0)]t^2 + \beta_0 t + y_0. \quad (3.23)$$

Figura 3.16



EJEMPLO 2 Las gráficas de la figura 3.16 muestran algunas posibilidades de curvas generadas por las ecuaciones (3.22) y (3.23) cuando los nodos son (0, 0) y (1, 0) y cuando las pendientes en estos nodos son 1 y −1, respectivamente. Las especificaciones de la pendiente en los extremos requieren sólo que $\alpha_0 = \beta_0$ y $\alpha_1 = -\beta_1$ porque las razones α_0/β_0 y α_1/β_1 dan las pendientes en los extremos izquierdo y derecho, respectivamente. ■

El procedimiento habitual con el que se determinan las curvas en un modo interactivo de graficación consiste en usar primero un dispositivo de entrada (un ratón o un *trackball*, por ejemplo) para fijar los nodos correctamente. Después se colocan los puntos guía para generar una primera aproximación a la curva deseada. Estos puntos pueden fijarse manualmente, pero la mayor parte de los sistemas de graficación permiten utilizar el dispositivo de entrada para dibujar manualmente la curva sobre la pantalla y seleccionar los nodos y puntos guías apropiados para esta curva.

Después podemos manipular los nodos y los puntos guía para colocarlos en una posición que produzca una curva satisfactoria desde el punto de vista estético. Como los cálculos son mínimos, generalmente podemos determinar la curva con tanta rapidez que el cambio resultante puede observarse de inmediato. Además, todos los datos necesarios para calcular las curvas están integrados en las coordenadas de los nodos y de los puntos guía; por tanto, el usuario del sistema no necesita un conocimiento analítico.

Los programas de graficación más comunes usan este tipo de sistema en las representaciones manuales, aunque en forma un poco distinta. Los polinomios cúbicos de Hermite se describen como polinomios de Bézier, los cuales incorporan un factor escalar de 3 cuando se calculan las derivadas en los extremos. Ello modifica las ecuaciones paramétricas y las transforma en

$$x(t) = [2(x_0 - x_1) + 3(\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - 3(\alpha_1 + 2\alpha_0)]t^2 + 3\alpha_0 t + x_0, \quad (3.24)$$

y en

$$y(t) = [2(y_0 - y_1) + 3(\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - 3(\beta_1 + 2\beta_0)]t^2 + 3\beta_0 t + y_0, \quad (3.25)$$

para $0 \le t \le 1$, pero este cambio no es advertido por el usuario del sistema.

Con el algoritmo 3.6 se construye un conjunto de curvas de Bézier basadas en las ecuaciones paramétricas en (3.24) y 3.25).

ALGORITMO 3.6

Curva de Bézier

Para construir las curvas de Bézier $C_0, \ldots, C_{n-1}$ en forma paramétrica, donde C_i está representada por

$$(x_i(t), y_i(t)) = (a_0^{(i)} + a_1^{(i)}t + a_2^{(i)}t^2 + a_3^{(i)}t^3, b_0^{(i)} + b_1^{(i)}t + b_2^{(i)}t^2 + b_3^{(i)}t^3),$$

para $0 \le t \le 1$, determinada por el extremo izquierdo (x_i, y_i), el punto guía de la izquierda (x_i^+, y_i^+), el extremo derecho (x_{i+1}, y_{i+1}) y el punto guía de la derecha (x_{i+1}^-, y_{i+1}^-) para cada $i = 0, 1, \ldots, n-1$:

ENTRADA n; $(x_0, y_0), \ldots, (x_n, y_n)$; $(x_0^+, y_0^+), \ldots, (x_{n-1}^+, y_{n-1}^+)$; $(x_1^-, y_1^-), \ldots, (x_n^-, y_n^-)$.

SALIDA coeficientes $\{a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, b_3^{(i)}$, para $0 \le i \le n-1\}$.

Paso 1 Para cada $i = 0, 1, \ldots, n-1$ hacer pasos 2 y 3.

Paso 2 Tome $a_0^{(i)} = x_i$;
$b_0^{(i)} = y_i$;
$a_1^{(i)} = 3(x_i^+ - x_i)$;
$b_1^{(i)} = 3(y_i^+ - y_i)$;
$a_2^{(i)} = 3(x_i + x_{i+1}^- - 2x_i^+)$;
$b_2^{(i)} = 3(y_i + y_{i+1}^- - 2y_i^+)$;
$a_3^{(i)} = x_{i+1} - x_i + 3x_i^+ - 3x_{i+1}^-$;
$b_3^{(i)} = y_{i+1} - y_i + 3y_i^+ - 3y_{i+1}^-$;

Paso 3 SALIDA $(a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)}, b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, b_3^{(i)})$.

Paso 4 PARAR. ■

Las curvas tridimensionales se generan en forma similar, especificando además los terceros componentes z_0 y z_1 para los nodos, así como $z_0 + \gamma_0$ y $z_1 - \gamma_1$, para los puntos guía. El problema más difícil relacionado con la representación de curvas tridimensionales es la pérdida de la tercera dimensión cuando proyectamos la curva en una pantalla de computadora bidimensional. Se usan varios métodos de proyección, pero este tema rebasa el ámbito de las gráficas por computadora. Si el lector desea una introducción a este tema y a las formas de modificar la técnica para obtener representaciones de superficies, le recomendamos cualquiera de las obras sobre métodos de gráficas por computadora, como el de [Hill, F].

CONJUNTO DE EJERCICIOS 3.5

1. Sean $(x_0, y_0) = (0, 0)$ y $(x_1, y_1) = (5, 2)$ los extremos de una curva. Use los puntos guía dados para construir las aproximaciones paramétricas cúbicas de Hermite $(x(t), y(t))$ a la curva y a la gráfica de las aproximaciones.
   a. (1, 1) y (6, 1) **b.** (0.5, 0.5) y (5.5, 1.5)
   c. (1, 1) y (6, 3) **d.** (2, 2) y (7, 0)
2. Repita el ejercicio 1 utilizando polinomios cúbicos de Bézier.
3. Construya y grafique los polinomios cúbicos de Bézier si se tienen los siguientes puntos y puntos guía.
   a. Punto (1, 1) con el punto guía (1.5, 1.25) al punto (6, 2) con el punto guía (7, 3)
   b. Punto (1, 1) con el punto guía (1.25, 1.5) al punto (6, 2) con el punto guía (5, 3)
   c. Punto (0, 0) con el punto guía (0.5, 0.5) al punto (4, 6) con el punto guía de entrada (3.5, 7) y el punto guía de salida (4.5, 5) al punto (6, 1) con el punto guía (7, 2)
   d. El punto (0, 0) con el punto guía (0.5, 0.5) al punto (2, 1) con el punto guía de entrada (3, 1) y el punto guía de salida (3, 1) al punto (4, 0) con el punto guía de entrada (5, 1) y el punto guía de salida (3, −1) al punto (6, −1) con el punto guía (6.5, −0.25).

4. Utilice los datos de la tabla siguiente y el algoritmo 3.6 para aproximar la forma de la letra *M*.

i	x_i	y_i	α_i	∂	α'_i	β'_i
0	3	6	3.3	6.5		
1	2	2	2.8	3.0	2.5	2.5
2	6	6	5.8	5.0	5.0	5.8
3	5	2	5.5	2.2	4.5	2.5
4	6.5	3			6.4	2.8

5. Suponga que hacemos pasar un polinomio cúbico de Bézier por (u_0, v_0) y (u_3, v_3) con los puntos guía (u_1, v_1) y (u_2, v_2) respectivamente.
   a. Deduzca las ecuaciones paramétricas de $u(t)$ y de $v(t)$ suponiendo que

   $$u(0) = u_0, \qquad u(1) = u_3, \qquad u'(0) = u_1 - u_0, \qquad u'(1) = u_3 - u_2.$$

   y que

   $$v(0) = v_0, \qquad v(1) = v_3, \qquad v'(0) = v_1 - v_0, \qquad v'(1) = v_3 - v_2.$$

   b. Sea $f(\frac{1}{3}i) = u_i$ para $i = 0, 1, 2, 3$ y sea $g(\frac{1}{3}i) = v_i$ para $i = 0, 1, 2, 3$. Demuestre que el polinomio de grado tres de Bernstein en t para f es $u(t)$ y que en t para g es $v(t)$. (Consulte el ejercicio 29 de la sección 3.1.)

3.6 Reseña de métodos y de software

En este capítulo estudiamos la aproximación de una función por medio de polinomios y de polinomios fragmentarios. Podemos especificar la función con una ecuación incluyendo puntos en el plano por donde pasa la gráfica de la función. Un conjunto de nodos $x_0, x_1, \ldots, x_n$ se da en cada caso, aunque tal vez se requiera más información; por ejemplo, el valor de algunas derivadas. El problema consiste en encontrar una función aproximante que cumpla con las condiciones especificadas por estos datos.

El polinomio interpolante $P(x)$ es el de menor grado que satisface, para una función f

$$P(x_i) = f(x_i) \qquad \text{para cada } i = 0, 1, \ldots, n$$

Aunque existe un único polinomio interpolante, éste puede adoptar muchísimas formas. La forma de Lagrange es la que más se emplea para interpolar las tablas, cuando n es pequeño y para derivar las fórmulas con las cuales aproximamos las derivadas e integrales. El método de Neville sirve para evaluar varios polinomios interpolantes en el mismo valor de x. Las formas de los polinomios de Newton son más adecuadas para los cálculos y también se usan mucho cuando se deducen fórmulas para resolver las ecuaciones diferenciales. Sin embargo, la interpolación polinómica presenta una debilidad intrínseca de oscilación, sobre todo si hay muchos nodos. En este caso disponemos de otros métodos cuya aplicación es más apropiada.

Los polinomios de Hermite interpolan una función y su derivada en los nodos, y pueden resultar sumamente precisos, pero requieren más información sobre la función a aproximar. También presentan la debilidad de oscilación cuando existen muchos nodos.

La forma más común de interpolación es polinómico-fragmentaria. Si se conocen los valores de la función y de la derivada, es recomendable aplicar la interpolación cúbica fragmentaria de Hermite. Este es el método preferido cuando se interpolan valores de una función que es la solución de una ecuación diferencial. Cuando sólo se conocen los valores de la función, podemos emplear la interpolación de trazadores cúbicos libres; esto hace que la segunda derivada del trazador sea cero en los extremos. Otros trazadores cúbicos requieren más datos. Por ejemplo, el trazador cúbico sujeto requiere que se conozcan los valores de la derivada de la función en los extremos del intervalo.

Existen otros métodos de interpolación que se utilizan con cierta frecuencia. La interpolación trigonométrica se emplea con grandes cantidades de datos cuando la función es de naturaleza periódica. En particular, se utiliza la transformada rápida de Fourier, la cual se explica en el capítulo 8. También se usa la interpolación por funciones racionales. Si se sospecha de la exactitud de los datos, podemos aplicar técnicas de alisamiento y recomendamos alguna forma de ajuste de los datos mediante los mínimos cuadrados. Cuando se realiza el ajuste con esta técnica, pueden utilizarse los polinomios, las funciones trigonométricas, las funciones racionales y los trazadores. Estos temas los estudiaremos en el capítulo 8.

Las rutinas de interpolación que se incluyen en la biblioteca IMSL se basan en el libro *A Practical Guide to Splines*, de Carl de Boor [Deb] y usan los trazadores cúbicos. La subrutina CSDEC se usa en la interpolación mediante trazadores cúbicos, y el usuario establece las condiciones finales; CSPER se usa en la interpolación mediante trazadores cúbicos con condiciones finales periódicas, y CSHER se emplea en la interpolación mediante polinomios fragmentarios cuasi-Hermite. La subrutina CSDEC incorpora los algoritmos 3.4 y 3.5. La subrutina CSINT se sirve de la condición no-un-nodo que se mencionó al final de la sección 3.4. También hay trazadores cúbicos que reducen al mínimo las oscilaciones o preservan la concavidad. También se incluyen métodos de interpolación bidimensional por medio de trazadores bicúbicos.

La biblioteca NAG /http://www.netlib.org contiene las subrutinas EO1AEF para la interpolación polinómica y la interpolación de Hermite, EO1BAF para la interpolación con trazadores cúbicos y EO1BEF para la interpolación cúbica fragmentaria de Hermite. La subrutina EO1ABF sirve para interpolar los datos en puntos igualmente espaciados. La rutina EO1AAF se aplica si se presentan los datos en puntos desigualmente espaciados. NAG contiene además subrutinas para interpolar las funciones de dos variables.

La biblioteca netlib contiene subrutinas cubspl.f en el paquete pppack para calcular el trazador cúbico con varias condiciones de extremos. En el paquete slatec, polint.f genera coeficientes de diferencia divididas de Newton para un conjunto discreto de puntos dato y en el paquete slatec/pchip hay varias rutinas para evaluar los polinomios fragmentarios de Hermite.

La función INTERP1 de MATLAB sirve para interpolar un conjunto discreto de puntos dato, usando la interpolación más cercana, interpolación lineal, interpolación trazado cúbico o interpolación cúbica. INTERP1 produce la evaluación polinomial en un conjunto discreto de puntos. POLYFIT, basado en la aproximación de mínimos cuadrados (véase la sección 8.1) puede usarse para encontrar una función interpolante de grado máximo n que cruza $n + 1$ puntos especificados. Con la función SPLINE podemos producir trazadores cúbicos. Maple construye un polinomio interpolante por medio del comando

```
> interp(X, Y, x);
```

donde `X` es la lista `[x[0],x[1],...,x[n]]`, `Y` es la lista `[f(x[0]], f(x[1]),..., f(x[n])` y x es la variable a utilizar.

El trazador cúbico natural también puede construirse con Maple. Primero introduzca

```
>readlib(spline);
```

para disponer del paquete. Con X y Y como en el párrafo anterior, el comando

```
>spline(X,Y,x,3):
```

construye el trazador cúbico natural que interpola `X=[x[0],...,x[n]]` y `Y=[y[0],...,y[n]]`, donde x es la variable y 3 indica el grado del trazador cúbico. También podemos crear trazadores lineales y cuadráticos.

Las obras de consulta general para los métodos que se estudiaron en este capítulo son los libros de Powell [Po] y de Davis [Da]. El primer trabajo que se dedicó a los trazadores es obra de Schoenberg [Scho]. Otros libros importantes sobre el tema son los de Schultz [Schul], De Boor [Deb] Dierex [Di] y Schumaker [Schum].

CAPÍTULO

4

Diferenciación e integración numéricas

. . .

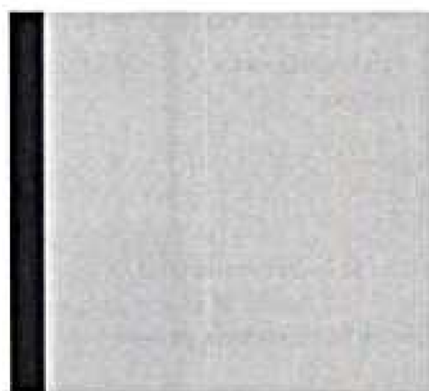
Se construye una hoja acanalada para techado, usando una máquina que comprime una hoja plana, de aluminio, y la transforma en una hoja cuya sección transversal tiene la forma de onda de la función seno.



Se necesita una hoja corrugada de 4 pies de largo cuyas ondas tienen una altura de 1 pulgada. desde la línea central, y cada onda tiene aproximadamente un periodo de 2π pulg. El problema de calcular la longitud de la primera hoja plana consiste en determinar la longitud de la onda dada por $f(x) = \text{sen}\, x$ de $x = 0$ pulg a $x = 48$ pulg. Por el cálculo sabemos que esta longitud es

$$L = \int_0^{48} \sqrt{1 + (f'(x))^2}\, dx = \int_0^{48} \sqrt{1 + (\cos x)^2}\, dx,$$

de modo que el problema consistirá en evaluar esta integral. Aunque la función seno es una de las más comunes en matemáticas, el cálculo de su longitud da origen a una integral elíptica de segunda clase, la cual no puede evaluarse con métodos normales. En este capítulo describiremos los métodos de aproximación de la solución a este tipo de problemas. Los abordaremos en particular en el ejercicio 21 de la sección 4.4 y en el ejercicio 10 de la sección 4.5.



En la introducción del capítulo 3 señalamos que una de las razones para que aproximemos un conjunto arbitrario de datos mediante polinomios algebraicos es que, dada una función continua cualquiera que esté definida en un intervalo cerrado, existirá un polinomio suficientemente cercano a la función en todos los puntos del intervalo. Por lo demás, las derivadas y las integrales de los polinomios se obtienen y se evalúan fácilmente. Por ello, no debería sorprendernos que la mayoría de los procedimientos para aproximar integrales y derivadas usen polinomios que aproximan la función.

4.1 Diferenciación numérica

La derivada de la función f en x_0 es

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Esta fórmula indica una manera obvia de generar una aproximación de $f'(x)$; basta calcular

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

para valores pequeños de h. Aunque esto parezca evidente, no es muy útil, debido a nuestro viejo enemigo, el error por redondeo. Sin embargo, ciertamente es un punto de partida.

Para aproximar $f'(x_0)$ supongamos primero que $x_0 \in (a, b)$, donde $f \in C^2[a, b]$, y que $x_1 = x_0 + h$ para alguna $h \neq 0$ que es lo bastante pequeña para asegurarnos de que $x_1 \in [a, b]$. Construimos el primer polinomio de Lagrange $P_{0,1}(x)$ para f determinada por x_0 y x_1 con su término de error:

$$\begin{aligned} f(x) &= P_{0,1}(x) + \frac{(x - x_0)(x - x_1)}{2!} f''(\xi(x)) \\ &= \frac{f(x_0)(x - x_0 - h)}{-h} + \frac{f(x_0 + h)(x - x_0)}{h} + \frac{(x - x_0)(x - x_0 - h)}{2} f''(\xi(x)), \end{aligned}$$

para alguna $\xi(x)$ en $[a, b]$. Al diferenciar obtenemos

$$\begin{aligned} f'(x) &= \frac{f(x_0 + h) - f(x_0)}{h} + D_x\left[\frac{(x - x_0)(x - x_0 - h)}{2} f''(\xi(x))\right] \\ &= \frac{f(x_0 + h) - f(x_0)}{h} + \frac{2(x - x_0) - h}{2} f''(\xi(x)) \\ &\quad + \frac{(x - x_0)(x - x_0 - h)}{2} D_x(f''(\xi(x))), \end{aligned}$$

de modo que

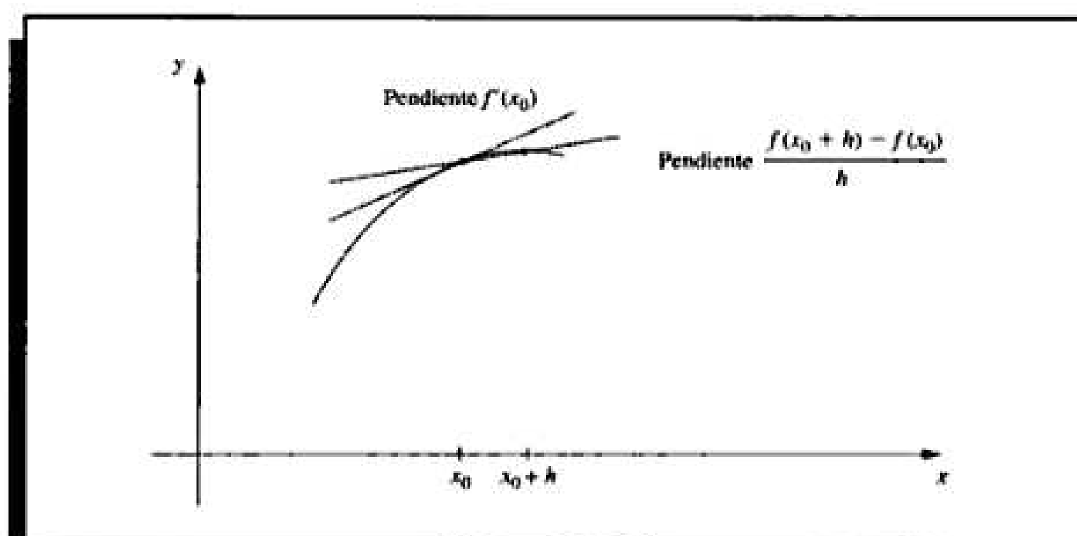
$$f'(x) \approx \frac{f(x_0 + h) - f(x_0)}{h}.$$

Un problema de esta fórmula radica en que carecemos de información sobre $D_x f''(\xi(x))$ por lo cual no podemos estimar el error de truncamiento. Pero cuando x es x_0 el coeficiente de $D_x f''(\xi(x))$ será cero y la fórmula se simplifica como sigue

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2} f''(\xi). \tag{4.1}$$

Para valores pequeños de h, podemos utilizar el cociente de la diferencia $[f(x_0 + h) - f(x_0)]/h$ para aproximar $f'(x_0)$ con un error acotado por $M|h|/2$, donde M es una cota en $|f''(x)|$ para $x \in [a, b]$. A esta fórmula se le llama **fórmula de la diferencia progresiva** si $h > 0$ (véase Fig. 4.1) y **fórmula de diferencia regresiva** si $h < 0$.

Figura 4.1



EJEMPLO 1 Sean $f(x) = \ln x$ y $x_0 = 1.8$. La fórmula de la diferencia progresiva

$$\frac{f(1.8 + h) - f(1.8)}{h}$$

sirve para aproximar $f'(1.8)$ con el error

$$\frac{|hf''(\xi)|}{2} = \frac{|h|}{2\xi^2} \leq \frac{|h|}{2(1.8)^2}, \qquad \text{donde } 1.8 < \xi < 1.8 + h.$$

Los resultados de la tabla 4.1 se producen cuando $h = 0.1$, 0.01 y 0.001.

Tabla 4.1

h	$f(1.8 + h)$	$\frac{f(1.8 + h) - f(1.8)}{h}$	$\frac{\lvert h \rvert}{2(1.8)^2}$
0.1	0.64185389	0.5406722	0.0154321
0.01	0.59332685	0.5540180	0.0015432
0.001	0.58834207	0.5554013	0.0001543

Puesto que $f'(x) = 1/x$, el valor exacto de $f'(1.8)$ es $0.55\overline{5}$ y las cotas de error son adecuadas. ■

Para obtener fórmulas de aproximación a la derivada más generales, supongamos que $\{x_0, x_1, \ldots, x_n\}$ son $(n + 1)$ números distintos en algún intervalo I y que $f \in C^{n+1}(I)$. Del teorema 3.3,

$$f(x) = \sum_{k=0}^{n} f(x_k)L_k(x) + \frac{(x - x_0)\cdots(x - x_n)}{(n + 1)!} f^{(n+1)}(\xi(x)),$$

para alguna $\xi(x)$ en I, donde $L_k(x)$ denota el k-ésimo polinomio de coeficiente de Lagrange para f en $x_0, x_1, \ldots, x_n$. Al diferenciar esta expresión obtenemos

$$f'(x) = \sum_{k=0}^{n} f(x_k)L_k'(x) + D_x\left[\frac{(x - x_0)\cdots(x - x_n)}{(n + 1!)}\right] f^{(n+1)}(\xi(x))$$

$$+ \frac{(x - x_0)\cdots(x - x_n)}{(n + 1)!} D_x[f^{(n+1)}(\xi(x))].$$

Una vez más tendremos un problema al estimar el error de truncamiento, a menos que x sea uno de los números x_j. En este caso, el término que contiene $D_x[f^{(n+1)}(\xi(x))]$ es cero, y entonces la fórmula queda así:

$$f'(x_j) = \sum_{k=0}^{n} f(x_k)L_k'(x_j) + \frac{f^{(n+1)}(\xi(x_j))}{(n + 1)!} \prod_{\substack{k=0\\k\neq j}}^{n} (x_j - x_k). \tag{4.2}$$

La ecuación (4.2) recibe el nombre de **fórmula de $(n + 1)$ puntos** para aproximar $f'(x_j)$.

En términos generales, la utilización de más puntos de evaluación en la ecuación (4.2) produce una mayor exactitud, aunque esto no conviene dada la cantidad de evaluaciones funcionales y el aumento en el error de redondeo. Las fórmulas más comunes son las que abarcan tres y cinco puntos de evaluación.

Primero derivamos alguna fórmula útil de tres puntos y consideramos los aspectos de sus errores. Puesto que

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \qquad \text{tenemos} \quad L_0'(x) = \frac{2x - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)}.$$

De manera análoga,

$$L_1'(x) = \frac{2x - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)} \quad \text{y} \quad L_2'(x) = \frac{2x - x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)}.$$

Por tanto, de acuerdo con la ecuación (4.2),

$$f'(x_j) = f(x_0)\left[\frac{2x_j - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)}\right] + f(x_1)\left[\frac{2x_j - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)}\right] \tag{4.3}$$

$$+ f(x_2)\left[\frac{2x_j - x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)}\right] + \frac{1}{6} f^{(3)}(\xi_j) \prod_{\substack{k=0\\k\neq j}}^{2} (x_j - x_k),$$

para cada $j = 0, 1, 2$, donde la notación ξ_j indica que este punto depende de x_j.

Las tres fórmulas de la ecuación (4.3) son de gran utilidad si los nodos son equidistantes, es decir, cuando

$$x_1 = x_0 + h \quad \text{y} \quad x_2 = x_0 + 2h, \qquad \text{para alguna } h \neq 0.$$

En el resto de esta sección supondremos que el espaciamiento de los nodos es igual.

Al utilizar la ecuación (4.3) con $x_j = x_0$, $x_1 = x_0 + h$ y con $x_2 = x_0 + 2h$ obtendremos

$$f'(x_0) = \frac{1}{h}\left[-\frac{3}{2}f(x_0) + 2f(x_1) - \frac{1}{2}f(x_2)\right] + \frac{h^2}{3}f^{(3)}(\xi_0).$$

Al hacer lo mismo con $x_j = x_1$, obtenemos

$$f'(x_1) = \frac{1}{h}\left[-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_2)\right] - \frac{h^2}{6}f^{(3)}(\xi_1),$$

y para $x_j = x_2$,

$$f'(x_2) = \frac{1}{h}\left[\frac{1}{2}f(x_0) - 2f(x_1) + \frac{3}{2}f(x_2)\right] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Puesto que $x_1 = x_0 + h$ y $x_2 = x_0 + 2h$, estas fórmulas también podemos expresarlas como

$$f'(x_0) = \frac{1}{h}\left[-\frac{3}{2}f(x_0) + 2f(x_0 + h) - \frac{1}{2}f(x_0 + 2h)\right] + \frac{h^2}{3}f^{(3)}(\xi_0),$$

$$f'(x_0 + h) = \frac{1}{h}\left[-\frac{1}{2}f(x_0) + \frac{1}{2}f(x_0 + 2h)\right] - \frac{h^2}{6}f^{(3)}(\xi_1) \quad \text{y}$$

$$f'(x_0 + 2h) = \frac{1}{h}\left[\frac{1}{2}f(x_0) - 2f(x_0 + h) + \frac{3}{2}f(x_0 + 2h)\right] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Por razones de comodidad, la sustitución de la variable $x_0 + h$ por x_0 se usa en la ecuación de en medio para transformar esta fórmula en una aproximación de $f'(x_0)$. Una sustitución semejante, $x_0 + 2h$, por x_0 se utiliza en la última ecuación. Esto nos da tres fórmulas para aproximar $f'(x_0)$:

$$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h)] + \frac{h^2}{3}f^{(3)}(\xi_0),$$

$$f'(x_0) = \frac{1}{2h}[-f(x_0 - h) + f(x_0 + h)] - \frac{h^2}{6}f^{(3)}(\xi_1) \quad \text{y}$$

$$f'(x_0) = \frac{1}{2h}[f(x_0 - 2h) - 4f(x_0 - h) + 3f(x_0)] + \frac{h^2}{3}f^{(3)}(\xi_2).$$

Finalmente nótese que, como podemos obtener la última ecuación a partir de la primera con sólo reemplazar h con $-h$, en realidad tenemos sólo dos fórmulas:

$$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h)] + \frac{h^2}{3}f^{(3)}(\xi_0), \qquad (4.4)$$

donde ξ_0 se encuentra entre x_0 y $x_0 + 2h$ y

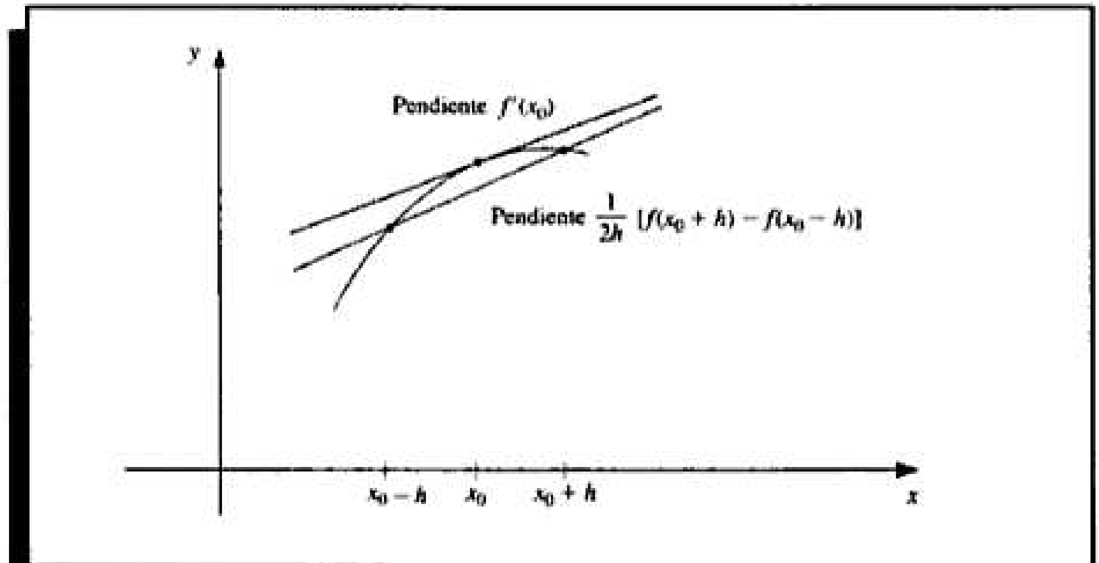
$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f^{(3)}(\xi_1), \qquad (4.5)$$

donde ξ_1 está entre $(x_0 - h)$ y $(x_0 + h)$.

Aunque los errores en (4.4) y (4.5) son $O(h^2)$, el error de la ecuación (4.5) es aproximadamente la mitad del error de la ecuación (4.4). Ello se debe a que en la ecuación (4.5) se emplean datos en ambos lados de x_0 y a que la ecuación (4.4) utiliza únicamente los de un lado. Asimismo, nótese que f debe evaluarse sólo en dos puntos en la ecuación (4.5), mientras que en la ecuación (4.4) se requieren tres evaluaciones. En la figura 4.2 se ilustra la aproximación producida con la ecuación (4.5).

En la ecuación (4.4) la aproximación es útil cerca de los extremos del intervalo ya que posiblemente no se tenga información de f fuera del intervalo.

Figura 4.2



Los métodos presentados en las ecuaciones (4.4) y (4.5) reciben el nombre de **fórmulas de tres puntos** (aunque el tercer punto $f(x_0)$ no aparezca en la ecuación (4.5)). Asimismo, llamadas **fórmulas de cinco puntos** en que se evalúa la función en dos puntos más, pero cuyo término de error tiene la forma $O(h^4)$. Una de esas fórmulas es

$$f'(x_0) = \frac{1}{12h}[f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)] + \frac{h^4}{30}f^{(5)}(\xi), \tag{4.6}$$

donde ξ está entre $x_0 - 2h$ y $x_0 + 2h$. Otra fórmula de cinco puntos de gran utilidad, sobre todo en lo relacionado con la interpolación de trazadores cúbicos sujetos de la sección 3.4, es la siguiente:

$$f'(x_0) = \frac{1}{12h}[-25f(x_0) + 48f(x_0 + h) - 36f(x_0 + 2h) \tag{4.7}$$

$$+ 16f(x_0 + 3h) - 3f(x_0 + 4h)] + \frac{h^4}{5}f^{(5)}(\xi),$$

donde ξ se encuentra entre x_0 y $x_0 + 4h$. Las aproximaciones del extremo izquierdo pueden obtenerse aplicando la fórmula con $h > 0$ y las aproximaciones del extremo derecho, con $h < 0$.

EJEMPLO 2 Los valores de $f(x) = xe^x$ están en la tabla 4.2.

Tabla 4.2

x	$f(x)$
1.8	10.889365
1.9	12.703199
2.0	14.778112
2.1	17.148957
2.2	19.855030

Puesto que $f'(x) = (x + 1)e^x$, tenemos $f'(2.0) = 22.167168$. Al aproximar $f'(2.0)$ mediante las fórmulas de tres y cinco puntos se obtienen los siguientes resultados.

Fórmulas de tres puntos

Usando (4.4) con $h = 0.1$: $\frac{1}{0.2}[-3f(2.0) + 4f(2.1) - f(2.2)] = 22.032310$,
Usando (4.4) con $h = -0.1$: $\frac{1}{-0.2}[-3f(2.0) + 4f(1.9) - f(1.8)] = 22.054525$,
Usando (4.5) con $h = 0.1$: $\frac{1}{0.2}[f(2.1) - f(1.9)] = 22.228790$,
Usando (4.5) con $h = 0.2$: $\frac{1}{0.4}[f(2.2) - f(1.8)] = 22.414163$,

Los errores en las fórmulas son aproximadamente

$$1.35 \times 10^{-1}, \quad 1.13 \times 10^{-1}, \quad -6.16 \times 10^{-2} \quad \text{y} \quad -2.47 \times 10^{-1},$$

respectivamente.

Fórmula de cinco puntos

Al utilizar (4.6) con $h = 0.1$ (la única fórmula de cinco puntos aplicable):

$$\frac{1}{1.2}[f(1.8) - 8f(1.9) + 8f(2.1) - f(2.2)] = 22.166996.$$

El error en esta fórmula es aproximadamente

$$1.69 \times 10^{-4}.$$

Está claro que la fórmula de cinco puntos da el mejor resultado. Observe además que el error de la ecuación (4.5) con $h = 0.1$ tiene aproximadamente la mitad de la magnitud del error que se genera al emplear la ecuación (4.4) con $h = 0.1$ o con $h = -0.1$. ■

También podemos derivar métodos para obtener las aproximaciones a derivadas de orden superior de una función, utilizando exclusivamente los valores tabulados de una función en varios puntos. Sin embargo, desde el punto de vista algebraico la derivación es tediosa y, por tanto, sólo describiremos un procedimiento representantivo.

Desarrollamos una función f en un tercer polinomio de Taylor alrededor de un punto x_0 y evaluemos en $x_0 + h$ y $x_0 - h$. Por tanto,

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \frac{1}{6}f'''(x_0)h^3 + \frac{1}{24}f^{(4)}(\xi_1)h^4$$

y

$$f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 - \frac{1}{6}f'''(x_0)h^3 + \frac{1}{24}f^{(4)}(\xi_{-1})h^4,$$

donde $x_0 - h < \xi_{-1} < x_0 < \xi_1 < x_0 + h$.

Si agregamos estas ecuaciones, el término $f'(x_0)$ se cancela y obtendremos

$$f(x_0 + h) + f(x_0 - h) = 2f(x_0) + f''(x_0)h^2 + \frac{1}{24}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})]h^4.$$

Al resolver $f''(x_0)$ en esta ecuación, obtenemos

$$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0 + h)] - \frac{h^2}{24}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})]. \quad (4.8)$$

Supongamos que $f^{(4)}$ es continua en $[x_0 - h, x_0 + h]$. Dado que $\frac{1}{2}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})]$ se encuentra entre $f^{(4)}(\xi_1)$ y $f^{(4)}(\xi_{-1})$, el teorema de valor intermedio implica que existe un número ξ entre ξ_1 y ξ_{-1}, y, por tanto, en $(x_0 - h, x_0 + h)$, con

$$f^{(4)}(\xi) = \frac{1}{2}[f^{(4)}(\xi_1) + f^{(4)}(\xi_{-1})].$$

Esto nos permite reescribir la ecuación (4.8) como

$$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0 + h)] - \frac{h^2}{12}f^{(4)}(\xi), \quad (4.9)$$

para alguna ξ, donde $x_0 - h < \xi < x_0 + h$.

EJEMPLO 3 Con los datos del ejemplo 2, para $f(x) = xe^x$ se puede usar la ecuación (4.9) para aproximar $f''(2.0)$. Puesto que $f''(x) = (x + 2)e^x$ el valor exacto será $f''(2.0) = 29.556224$. Al emplear (4.9) con $h = 0.1$ obtenemos

$$f''(2.0) \approx \frac{1}{0.01}[f(1.9) - 2f(2.0) + f(2.1)] = 29.593200,$$

al emplear (4.9) con $h = 0.2$ obtenemos

$$f''(2.0) \approx \frac{1}{0.04}[f(1.8) - 2f(2.0) + f(2.2)] = 29.704275.$$

Los errores son aproximadamente -3.70×10^{-2} y -1.48×10^{-1}, respectivamente. ∎

Un punto muy importante en el estudio de la diferenciación numérica es el efecto que tiene el error de redondeo en la aproximación. Examinemos más detenidamente la ecuación (4.5):

$$f'(x_0) = \frac{1}{2h}[f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6}f^{(3)}(\xi_1).$$

Supongamos que al evaluar $f(x_0 + h)$ y $f(x_0 - h)$ descubriremos los errores de redondeo $e(x_0 + h)$ y $e(x_0 - h)$. Entonces los valores calculados $\tilde{f}(x_0 + h)$ y $\tilde{f}(x_0 - h)$ se relacionan con los valores verdaderos $f(x_0 + h)$ y $f(x_0 - h)$ por medio de las fórmulas:

$$f(x_0 + h) = \tilde{f}(x_0 + h) + e(x_0 + h)$$

y

$$f(x_0 - h) = \tilde{f}(x_0 - h) + e(x_0 - h)$$

El error total de la aproximación,

$$f'(x_0) - \frac{\tilde{f}(x_0 + h) - \tilde{f}(x_0 - h)}{2h} = \frac{e(x_0 + h) - e(x_0 - h)}{2h} - \frac{h^2}{6} f^{(3)}(\xi_1),$$

tendrá una parte debida al error de redondeo y otra al error de truncamiento. Si suponemos que los errores de redondeo $e(x_0 \pm h)$ están acotados por algún número $\varepsilon > 0$ y que la tercera derivada de f está acotada por un número $M > 0$, entonces

$$\left| f'(x_0) - \frac{\tilde{f}(x_0 + h) - \tilde{f}(x_0 - h)}{2h} \right| \le \frac{\varepsilon}{h} + \frac{h^2}{6} M.$$

Si queremos reducir el error de truncamiento, $h^2M/6$, debemos reducir h. Pero al reducir h, el error de redondeo ε/h crece. Así pues, en la práctica rara vez conviene que h sea muy pequeño, porque el error de redondeo predominará en los cálculos.

EJEMPLO 4 Use los valores de la tabla 4.3 para aproximar $f'(0.900)$, donde $f(x) = \operatorname{sen} x$. El valor verdadero es $\cos 0.900 = 0.62161$.

Tabla 4.3

x	$\operatorname{sen} x$	x	$\operatorname{sen} x$
0.800	0.71736	0.901	0.78395
0.850	0.75128	0.902	0.78457
0.880	0.77074	0.905	0.78643
0.890	0.77707	0.910	0.78950
0.895	0.78021	0.920	0.79560
0.898	0.78208	0.950	0.81342
0.899	0.78270	1.000	0.84147

Al aplicar la fórmula

$$f'(0.900) \approx \frac{f(0.900 + h) - f(0.900 - h)}{2h},$$

con diferentes valores de h, obtenemos las aproximaciones de la tabla 4.4.

La elección óptima de h parece encontrarse entre 0.005 y 0.05. Si analizamos un poco el término de error,

$$e(h) = \frac{\varepsilon}{h} + \frac{h^2}{6} M,$$

Tabla 4.4

h	Aproximación a $f'(0.900)$	Error
0.001	0.62500	0.00339
0.002	0.62250	0.00089
0.005	0.62200	0.00039
0.010	0.62150	−0.00011
0.020	0.62150	−0.00011
0.050	0.62140	−0.00021
0.100	0.62055	−0.00106

podremos utilizar el cálculo para verificar (véase el ejercicio 23) que un mínimo de e ocurre en $h = \sqrt[3]{3\varepsilon / M}$, donde

$$M = \max_{x\in[0.800,1.00]} |f'''(x)| = \max_{x\in[0.800,1.00]} |\cos x| = \cos 0.8 \approx 0.69671.$$

Como los valores de f se dan en cinco cifras decimales, es razonable suponer que el error de redondeo está acotado por $\varepsilon = 0.000005$. Por tanto, la elección óptima de h es aproximadamente

$$h = \sqrt[3]{\frac{3(0.000005)}{0.69671}} \approx 0.028,$$

lo cual es compatible con los resultados de la tabla 4.4.

En la práctica, no es posible calcular un valor óptimo de h que nos sirva para aproximar la derivada, ya que no conocemos la tercera derivada de la función. Pero no debemos olvidar que con la reducción del tamaño del paso no siempre mejoraremos la aproximación. ■

Aunque sólo hemos analizado los problemas del error de redondeo que se presentan con la ecuación de la fórmula de tres puntos (4.5), se presentan problemas semejantes con todas las fórmulas de diferenciación. Ello se debe a la necesidad de dividir una potencia de h. Como comprobamos en la sección 1.2 (véase, en particular, el ejemplo 3), la división entre números pequeños tiende a exagerar el error del redondeo, por lo que debería evitarse en lo posible. En el caso de la diferenciación numérica, es imposible evitar por completo el problema, aunque sí lo atenuamos con los métodos de orden superior.

Recuerde que, como método de aproximación, la diferenciación numérica es *inestable* porque los valores pequeños de h necesarios para disminuir el error de truncamiento, también hacen crecer el error de redondeo. Esta es la primera clase de métodos inestables que hemos encontrado y, en lo posible, deberíamos omitirlos. Sin embargo, además de emplearse en los cálculos, las fórmulas que hemos derivado son necesarias para aproximar las soluciones de las ecuaciones diferenciales ordinarias y parciales.

CONJUNTO DE EJERCICIOS 4.1

1. Use las fórmulas de diferencia progresiva y de diferencia regresiva para determinar las aproximaciones con que se completarán las siguientes tablas.

a.

x	$f(x)$	$f'(x)$
0.5	0.4794	
0.6	0.5646	
0.7	0.6442	

b.

x	$f(x)$	$f'(x)$
0.0	0.00000	
0.2	0.74140	
0.4	1.3718	

2. Los datos del ejercicio 1 se tomaron de las siguientes funciones. Calcule los errores reales del ejercicio 1 y obtenga las cotas de error por medio de las fórmulas de error.

a. $f(x) = \text{sen } x$ b. $f(x) = e^x - 2x^2 + 3x - 1$

3. Use la fórmula de tres puntos más conveniente para determinar las aproximaciones con que se completarán las siguientes tablas.

a.

x	$f(x)$	$f'(x)$
1.1	9.025013	
1.2	11.02318	
1.3	13.46374	
1.4	16.44465	

b.

x	$f(x)$	$f'(x)$
8.1	16.94410	
8.3	17.56492	
8.5	18.19056	
8.7	18.82091	

c.

x	$f(x)$	$f'(x)$
2.9	−4.827866	
3.0	−4.240058	
3.1	−3.496909	
3.2	−2.596792	

d.

x	$f(x)$	$f'(x)$
2.0	3.6887983	
2.1	3.6905701	
2.2	3.6688192	
2.3	3.6245909	

4. Los datos del ejercicio 3 se tomaron de las siguientes funciones. Calcule los errores reales del ejercicio 3 y obtenga las cotas de error por medio de las fórmulas de error.

a. $f(x) = e^{2x}$ b. $f(x) = x \ln x$

c. $f(x) = x \cos x - x^2 \text{ sen } x$ d. $f(x) = 2(\ln x)^2 + 3 \text{ sen } x$

5. Use la fórmula más precisa posible de esta sección para determinar las aproximaciones con que se completarán las siguientes tablas.

a.

x	$f(x)$	$f'(x)$
2.1	−1.709847	
2.2	−1.373823	
2.3	−1.119214	
2.4	−0.9160143	
2.5	−0.7470223	
2.6	−0.6015966	

b.

x	$f(x)$	$f'(x)$
−3.0	9.367879	
−2.8	8.233241	
−2.6	7.180350	
−2.4	6.209329	
−2.2	5.320305	
−2.0	4.513417	

6. Los datos del ejercicio 5 se tomaron de las funciones dadas. Calcule los errores reales del ejercicio 5 y obtenga las cotas de error usando las fórmulas de error y Maple.

a. $f(x) = \tan x$ b. $f(x) = e^{x/3} + x^2$

7. Use los siguientes datos y el hecho de que las primeras cinco derivadas de f estaban acotadas en [1,5] por 2, 3, 6, 12 y 23, respectivamente, para aproximar $f'(3)$ con la mayor exactitud posible. Obtenga una cota del error.

x	1	2	3	4	5
$f(x)$	2.4142	2.6734	2.8974	3.0976	3.2804

8. Repita el ejercicio 7, pero suponga que la tercera derivada de f está acotada en [1, 5] por 4.

9. Repita el ejercicio 1 usando la aritmética de redondeo a cuatro dígitos, y después compare los errores con los del ejercicio 2.

10. Repita el ejercicio 3 usando la aritmética de corte de cuatro dígitos, y después compare los errores con los del ejercicio 4.

11. Repita el ejercicio 5 usando la aritmética de redondeo a cuatro dígitos, y después compare los errores con los del ejercicio 6.

12. Estudie la siguiente tabla de datos.

x	0.2	0.4	0.6	0.8	1.0
$f(x)$	0.9798652	0.9177710	0.808038	0.6386093	0.3843735

a. Aplique las fórmulas adecuadas para aproximar $f'(0.4)$ y $f''(0.4)$.

b. Aplique las fórmulas adecuadas para aproximar $f'(0.6)$ y $f''(0.6)$.

13. Sea $f(x) = \cos \pi x$. Use la ecuación (4.9) y los valores de $f(x)$ en $x = 0.25$, 0.5 y 0.75 para aproximar $f''(0.5)$. Compare este resultado con el valor exacto y con la aproximación que se obtuvo en el ejercicio 11 de la sección 3.4. Explique por qué este método resulta tan exacto en este problema. Obtenga una cota de error.

14. Sea $f(x) = 3xe^x - \cos x$. Use los siguientes datos y la ecuación (4.9) para aproximar $f''(1.3)$ con $h = 0.1$ y con $h = 0.01$.

x	1.20	1.29	1.30	1.31	1.40
$f(x)$	11.59006	13.78176	14.04276	14.30741	16.86187

Compare los resultados con $f''(1.3)$.

15. Examine detenidamente la siguiente tabla de datos:

x	0.2	0.4	0.6	0.8	1.0
$f(x)$	0.9798652	0.9177710	0.8080348	0.6386093	0.3843735

a. Use la ecuación (4.7) para aproximar $f'(0.2)$.

b. Use la ecuación (4.7) para aproximar $f'(1.0)$.

c. Use la ecuación (4.6) para aproximar $f'(0.6)$.

16. Derive una fórmula de cinco puntos $O(h^4)$ para aproximar $f'(x_0)$ que utilice $f(x_0 - h)$, $f(x_0)$, $f(x_0 + h)$, $f(x_0 + 2h)$ y $f(x_0 + 3h)$. [*Sugerencia*: considere la expresión $Af(x_0 - h) + Bf(x_0 + h) + Cf(x_0 + 2h) + Df(x_0 + 3h)$. Desarrollo en términos del quinto polinomio de Taylor y seleccione A, C y D apropiadamente.]

17. Aplique la fórmula derivada en el ejercicio 16 y los datos del ejercicio 15 para aproximar $f'(0.4)$ y $f'(0.8)$.

18. **a.** Analice los errores de redondeo, como en el ejemplo 4, para la fórmula

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2} f''(\xi_0).$$

b. Encuentre un valor óptimo de $h > 0$ para la función dada en el ejemplo 2.

19. En el ejercicio 7 de la sección 3.3 se incluyeron datos que describen un automóvil que recorre una carretera recta. En el problema se podía predecir su posición y velocidad cuando $t = 10$ s. Use los siguientes tiempos y posiciones para predecir la velocidad del automóvil en cada momento incluido en la tabla.

Tiempo	0	3	5	8	10	13
Distancia	0	225	383	623	742	993

20. En un circuito con un voltaje impreso $\varepsilon(t)$ y una inductancia L, la primera ley de Kirchhoff nos da la siguiente relación

$$\varepsilon(t) = L\frac{di}{dt} + Ri,$$

donde R es la resistencia del circuito e i es la corriente. Suponga que medimos la corriente con varios valores de t y obtenemos:

t	1.00	1.01	1.02	1.03	1.0
i	3.10	3.12	3.14	3.18	3.24

donde t se mide en segundos, i se da en amperes, la inductancia L es una constante de 0.98 henries y la resistencia es de 0.142 ohms. Aproxime el voltaje $\varepsilon(t)$ en los valores $t = 1.00$, 1.01, 1.02, 1.03 y 1.04.

21. Los estudiantes de cálculo saben que la derivada de una función f en x puede definirse como

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

Escoja su función favorita f, un número no cero x y utilice una computadora o calculadora. Genere las aproximaciones $f_n'(x)$ por $f'(x)$ para

$$f_n'(x) = \frac{f(x + 10^{-n}) - f(x)}{10^{-n}},$$

para $n = 1, 2, \ldots, 20$ y describa lo que sucede.

22. Deduzca un método para aproximar $f'''(x_0)$, cuyo término de error sea del orden h^2, desarrollando para ello la función f en un cuarto polinomio de Taylor alrededor de x_0 y evaluando en $x_0 \pm h$ y en $x_0 \pm 2h$.

23. Examine detenidamente la función

$$e(h) = \frac{\varepsilon}{h} + \frac{h^2}{6}M,$$

donde M es una cota de la tercera derivada de una función. Demuestre que $e(h)$ tiene un mínimo en $\sqrt[3]{3\varepsilon/M}$.

4.2 Extrapolación de Richardson

La extrapolación de Richardson sirve para generar resultados de gran exactitud cuando se usan fórmulas de bajo orden. Aunque el nombre dado al método se refiere a un trabajo realizado por L. F. Richardson y J. A. Gaunt [RG] en 1927, la idea en que se basa esta técnica es mucho más antigua. En [Joy] usted encontrará un artículo interesante sobre la historia y la aplicación de la extrapolación.

La extrapolación puede aplicarse siempre que sepamos que el método de aproximación tiene un término de error de una forma previsible; la forma se basa en un parámetro, que generalmente es el tamaño de paso h. Supongamos que, para cada número $h \neq 0$ te-

nemos una fórmula $N(h)$ que aproxima un valor desconocido M y que el error de truncamiento que supone la aproximación presenta la forma

$$M - N(h) = K_1h + K_2h^2 + K_3h^3 + \cdots,$$

para algún conjunto de constantes desconocidas $K_1, K_2, K_3, \ldots$

Dado que el error de truncamiento es $O(h)$, podríamos esperar, por ejemplo, que

$$M - N(0.1) \approx 0.1K_1, \quad M - N(0.01) \approx 0.01K_1,$$

y, en general, que $M - N(h) \approx K_1h$, salvo que haya una gran variación de magnitud entre las constantes $K_1, K_2, K_3, \ldots$

La extrapolación tiene por objeto encontrar un modo fácil de combinar las aproximaciones bastante imprecisas $O(h)$ en forma apropiada para producir fórmulas con un error de truncamiento de orden superior. Supongamos, por ejemplo, que pudiéramos combinar las $N(h)$ fórmulas así como producir una fórmula de aproximación $O(h^2)$, $\hat{N}(h)$, para M con

$$M - \hat{N}(h) = \hat{K}_2h^2 + \hat{K}_3h^3 + \cdots,$$

una vez más, para un conjunto desconocido de constantes $\hat{K}_1, \hat{K}_2, \ldots$ Entonces podríamos tener

$$M - \hat{N}(0.1) \approx 0.01\hat{K}_2, \qquad M - \hat{N}(0.001) \approx 0.0001\hat{K}_2,$$

y así sucesivamente. Si las constantes K_1 y $\hat{K}_2$ son aproximadamente de la misma magnitud entonces las aproximaciones $\hat{N}(h)$ serán mucho mejores que las aproximaciones $N(h)$ correspondientes. La extrapolación continúa al combinar las aproximaciones $\hat{N}(h)$ en forma tal que produzcan fórmulas con un error de truncamiento $O(h^3)$ y así sucesivamente.

Para ver concretamente cómo podemos generar estas fórmulas de orden superior, tomemos el caso de la fórmula con que se aproxima M de la forma

$$M = N(h) + K_1h + K_2h^2 + K_3h^3 + \cdots. \tag{4.10}$$

Suponemos que la fórmula se aplica a cualquier h positivo, por lo cual consideraremos el resultado cuando reemplacemos el parámetro h por la mitad de su valor. Tenemos entonces la fórmula

$$M = N\left(\frac{h}{2}\right) + K_1\frac{h}{2} + K_2\frac{h^2}{4} + K_3\frac{h^3}{8} + \cdots.$$

Al restar la ecuación (4.10) dos veces de esta ecuación, eliminamos el término que contiene K_1 y obtenemos

$$M = \left[N\left(\frac{h}{2}\right) + \left(N\left(\frac{h}{2}\right) - N(h)\right)\right] + K_2\left(\frac{h^2}{2} - h^2\right) + K_3\left(\frac{h^3}{4} - h^3\right) + \cdots.$$

Con el fin de facilitar la explicación, definimos $N_1(h) \equiv N(h)$ y

$$N_2(h) = N_1\left(\frac{h}{2}\right) + \left[N_1\left(\frac{h}{2}\right) - N_1(h)\right].$$

Tenemos entonces la fórmula $O(h^2)$ con que aproximamos M:

$$M = N_2(h) - \frac{K_2}{2}h^2 - \frac{3K_3}{4}h^3 - \cdots. \tag{4.11}$$

Si ahora reemplazamos h por $h/2$ en esta fórmula, obtendremos

$$M = N_2\left(\frac{h}{2}\right) - \frac{K_2}{8}h^2 - \frac{3K_3}{32}h_3 - \cdots . \tag{4.12}$$

Podemos combinar la fórmula anterior con la ecuación (4.11) para suprimir el término h^2. En concreto, al restar la ecuación (4.11) 4 veces de la ecuación (4.12) obtenemos

$$3M = 4N_2\left(\frac{h}{2}\right) - N_2(h) + \frac{3K_3}{8}h^3 + \cdots ,$$

y dividir entre 3 nos da la fórmula $O(h^3)$ para aproximar M:

$$M = \left[N_2\left(\frac{h}{2}\right) + \frac{N_2(h/2) - N_2(h)}{3}\right] + \frac{K_3}{8}h^3 + \cdots .$$

Al definir

$$N_3(h) = N_2\left(\frac{h}{2}\right) + \frac{N_2(h/2) - N_2(h)}{3},$$

tenemos la fórmula $O(h^3)$:

$$M = N_3(h) + \frac{K_3}{8}h^3 + \cdots .$$

El proceso continúa al construir aproximación $O(h^4)$

$$N_4(h) = N_3\left(\frac{h}{2}\right) + \frac{N_3(h/2) - N_3(h)}{7},$$

la aproximación $O(h^5)$

$$N_5(h) = N_4\left(\frac{h}{2}\right) + \frac{N_4(h/2) - N_4(h)}{15},$$

y así sucesivamente. En general, si M puede reescribirse en la forma

$$M = N(h) + \sum_{j=1}^{m-1} K_j h^j + O(h^m), \tag{4.13}$$

entonces para cada $j = 2, 3, \ldots, m$, tendremos una aproximación $O(h^j)$ de la forma

$$N_j(h) = N_{j-1}\left(\frac{h}{2}\right) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{2^{j-1} - 1}. \tag{4.14}$$

Las aproximaciones anteriores se generan por renglones o filas en el orden indicado por los valores o entradas numeradas de la tabla 4.5. Esto se hace para aprovechar al máximo las fórmulas de orden superior.

La extrapolación puede aplicarse siempre que el error de truncamiento de una fórmula presente la forma

$$\sum_{j=1}^{m-1} K_j h^{\alpha_j} + O(h^{\alpha_m}),$$

Tabla 4.5

$O(h)$	$O(h^2)$	$O(h^3)$	$O(h^4)$
1: $N_1(h) \equiv N(h)$			
2: $N_1(\frac{h}{2}) \equiv N(\frac{h}{2})$	3: $N_2(h)$		
4: $N_1(\frac{h}{4}) \equiv N(\frac{h}{4})$	5: $N_2(\frac{h}{2})$	6: $N_3(h)$	
7: $N_1(\frac{h}{8}) \equiv N(\frac{h}{8})$	8: $N_2(\frac{h}{4})$	9: $N_3(\frac{h}{2})$	10: $N_4(h)$

para un conjunto de constantes K_j y cuando $\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_m$. En el siguiente ejemplo tenemos $\alpha_j = 2j$.

EJEMPLO 1 En la ecuación (4.5), la fórmula de la diferencia centrada para aproximar $f'(x_0)$ puede expresarse mediante una fórmula de error:

$$f'(x_0) = \frac{1}{2h}[f(x_0+h) - f(x_0-h)] - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(x_0) - \cdots .$$

Puesto que esta fórmula de error contiene exclusivamente potencias pares de h, la extrapolación es más eficaz de lo que sugiere la explicación inicial. En este caso tenemos la aproximación $O(h^2)$

$$f'(x_0) = N_1(h) - \frac{h^2}{6}f'''(x_0) - \frac{h^4}{120}f^{(5)}(x_0) - \cdots , \qquad (4.15)$$

donde

$$N_1(h) \equiv N(h) = \frac{1}{2h}[f(x_0+h) - f(x_0-h)].$$

Al reemplazar h por $h/2$ en esta fórmula, obtenemos la aproximación

$$f'(x_0) = N_1\left(\frac{h}{2}\right) - \frac{h^2}{24}f'''(x_0) - \frac{h^4}{1920}f^{(5)}(x_0) - \cdots .$$

Al restar la ecuación (4.15) 4 veces de esta ecuación suprimimos el término $O(h^2)$ que contiene $f'''(x_0)$ y obtenemos

$$3f'(x_0) = 4N_1\left(\frac{h}{2}\right) - N_1(h) + \frac{h^4}{160}f^{(5)}(x_0) + \cdots .$$

Al dividir entre 3 obtenemos la fórmula $O(h^4)$

$$f'(x_0) = N_2(h) + \frac{h^4}{480}f^{(5)}(x_0) + \cdots ,$$

donde

$$N_2(h) = N_1\left(\frac{h}{2}\right) + \frac{N_1(h/2) - N_1(h)}{3}.$$

Al continuar con este procedimiento obtenemos, para cada $j = 2, 3, \ldots,$ la aproximación $O(h^{2j})$

$$N_j(h) = N_{j-1}\left(\frac{h}{2}\right) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{4^{j-1} - 1}.$$

Nótese que el denominador del cociente es ahora $4^{j-1} - 1$ en vez de $2^{j-1} - 1$ porque ahora estamos eliminando potencias de h^2 y no potencias de h. Dado que $(h/2)^2 = h^2/4$, los multiplicadores con que eliminamos las potencias de h^2 son potencias de 4 y no de 2.

Supongamos que $x_0 = 2.0$, $h = 0.2$ y $f(x) = xe^x$. Entonces

$$N_1(0.2) = N(0.2) \frac{1}{0.4} [f(2.2) - f(1.8)] = 22.414160,$$

$$N_1(0.1) = N(0.1) = 22.228786,$$

y

$$N_1(0.05) = N(0.05) = 22.182564.$$

La tabla con que extrapolamos los datos anteriores aparece en la tabla 4.6. El valor exacto de $f'(x) = xe^x + e^x$ a $x_0 = 2.0$ con seis decimales es 22.167168 de modo que todos los dígitos de $N_3(0.2)$ son exactos, aunque la mejor aproximación original, $N_1(0.5)$, sólo tenía una cifra decimal de precisión. ■

Tabla 4.6

$N_1(0.2) = 22.414160$		
$N_1(0.1) = 22.228786$	$N_2(0.2) = N_1(0.1) + \frac{N_1(0.1) - N_1(0.2)}{3}$ $= 22.166995$	
$N_1(0.05) = 22.182564$	$N_2(0.1) = N_1(0.05) + \frac{N_1(0.05) - N_1(0.1)}{3}$ $= 22.167157$	$N_3(0.2) = N_2(0.1) + \frac{N_2(0.1) - N_2(0.2)}{15}$, $= 22.167168$

En la tabla de extrapolación, las columnas después de la primera se obtienen con un simple proceso de prorrateo; por ello el método puede producir aproximaciones de orden elevado con un mínimo de cálculos y de error de redondeo. Sin embargo, conforme k aumenta, el error de redondeo en $N_1(h/2^k)$ generalmente se incrementará a causa de que la inestabilidad de la diferenciación numérica está relacionada con el tamaño de paso $h/2^k$.

En la sección 4.1 explicamos los métodos de tres y cinco puntos con los cuales se aproxima $f'(x_0)$ cuando se conocen varios valores funcionales de f. Los métodos de tres puntos se derivaron diferenciando un polinomio interpolante de Lagrange para f. Los métodos de cinco puntos se pueden obtener en una forma parecida, pero la derivación resulta tediosa, por lo que podemos utilizar la extrapolación para derivar más fácilmente esas fórmulas.

Supongamos que desarrollamos la función f en el cuarto polinomio de Taylor alrededor de x_0. Entonces

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \frac{1}{6} f'''(x_0)(x - x_0)^3$$

$$+ \frac{1}{24} f^{(4)}(x_0)(x - x_0)^4 + \frac{1}{120} f^{(5)}(\xi)(x - x_0)^5,$$

para algún número ξ entre x y x_0. Al evaluar f en $x_0 + h$ y $x_0 - h$ obtenemos

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2} f''(x_0)h^2 + \frac{1}{6} f'''(x_0)h^3 \tag{4.16}$$

$$+ \frac{1}{24} f^{(4)}(x_0)h^4 + \frac{1}{120} f^{(5)}(\xi_1)h^5$$

y

$$f(x_0 - h) = f(x_0) - f'(x_0)h + \frac{1}{2} f''(x_0)h^2 - \frac{1}{6} f'''(x_0)h^3 \tag{4.17}$$

$$+ \frac{1}{24} f^{(4)}(x_0)h^4 - \frac{1}{120} f^{(5)}(\xi_2)h^5,$$

donde $x_0 - h < \xi_2 < x_0 < \xi_1 < x_0 + h$. Al restar la ecuación (4.17) a la ecuación (4.16) obtenemos

$$f(x_0 + h) - f(x_0 - h) = 2h f'(x_0) + \frac{h^3}{3} f'''(x_0) + \frac{h^5}{120} [f^{(5)}(\xi_1) + f^{(5)}(\xi_2)]. \tag{4.18}$$

Si $f^{(5)}$ es continua en $[x_0 - h, x_0 + h]$, el teorema del valor intermedio implica que existe un número $\tilde{\xi}$ en $(x_0 - h, x_0 + h)$ con

$$f^{(5)}(\tilde{\xi}) = \frac{1}{2}[f^{(5)}(\xi_1) + f^{(5)}(\xi_2)].$$

En consecuencia, podemos despejar $f'(x_0)$ en la ecuación (4.18) para obtener la aproximación $O(h^2)$

$$f'(x_0) = \frac{1}{2h} [f(x_0 + h) - f(x_0 - h)] - \frac{h^2}{6} f'''(x_0) - \frac{h^4}{120} f^{(5)}(\tilde{\xi}). \tag{4.19}$$

Aunque la aproximación de la ecuación (4.19) es la misma que la de la fórmula de tres puntos en la ecuación (4.5), el punto desconocido de la evaluación ocurre ahora en $f^{(5)}$, y no en f'''. Esto se aprovecha en la extrapolación, primero reemplazando h en la ecuación (4.19) por $2h$ para producir la nueva fórmula

$$f'(x_0) = \frac{1}{4h} [f(x_0 + 2h) - f(x_0 - 2h)] - \frac{4h^2}{6} f'''(x_0) - \frac{16h^4}{120} f^{(5)}(\hat{\xi}), \tag{4.20}$$

donde $\hat{\xi}$ está entre $x_0 - 2h$ y $x_0 + 2h$.

Al multiplicar la ecuación (4.19) por 4 y al restar la ecuación (4.20), obtenemos

$$3f'(x_0) = \frac{2}{h} [f(x_0 + h) - f(x_0 - h)] - \frac{1}{4h} [f(x_0 + 2h) - f(x_0 - 2h)]$$

$$- \frac{h^4}{30} f^{(5)}(\tilde{\xi}) + \frac{2h^4}{15} f^{(5)}(\hat{\xi}).$$

Si $f^{(5)}$ es continua en $[x_0 - 2h, x_0 + 2h]$, podemos aplicar otro método para demostrar que $f^{(5)}(\tilde{\xi})$ y $f^{(5)}(\hat{\xi})$ puede reemplazarse con un valor común $f^{(5)}(\xi)$. Al utilizar este resultado y al dividir entre 3 obtenemos la fórmula de cinco puntos

$$f'(x_0) = \frac{1}{12h} [f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)] + \frac{h^4}{30} f^{(5)}(\xi),$$

que es la que aparece en la ecuación (4.6). Otras fórmulas de las derivadas primera y de orden superior pueden obtenerse en forma parecida; algunas de ellas vienen en los ejercicios.

A lo largo del texto utilizamos el método de extrapolación. Las aplicaciones más importantes son la aproximación de integrales, en la sección 4.5, y la determinación de soluciones aproximadas de las ecuaciones diferenciales, en la sección 5.8.

CONJUNTO DE EJERCICIOS 4.2

1. Use la técnica de extrapolación que se describe en el ejemplo 1 para determinar $N_3(h)$ una aproximación de $f'(x_0)$, para las siguientes funciones y tamaños de paso.
   a. $f(x) = \ln x$, $x_0 = 1.0$, $h = 0.4$
   b. $f(x) = x + e^x$, $x_0 = 0.0$, $h = 0.4$
   c. $f(x) = 2^x \operatorname{sen} x$, $x_0 = 1.05$, $h = 0.4$
   d. $f(x) = x^3 \cos x$, $x_0 = 2.3$, $h = 0.4$
2. Agregue otra línea a la tabla de extrapolación del ejercicio 1 para obtener la aproximación $N_4(h)$.
3. Repita el ejercicio 1 usando una aritmética de redondeo a cuatro dígitos.
4. Repita el ejercicio 2 usando una aritmética de redondeo a cuatro dígitos.
5. En los siguientes datos se dan las aproximaciones de la integral

$$M = \int_0^{\pi} \operatorname{sen} x\, dx.$$

$$N_1(h) = 1.570796, \quad N_1\left(\frac{h}{2}\right) = 1.896119, \quad N_1\left(\frac{h}{4}\right) = 1.974232, \quad N_1\left(\frac{h}{8}\right) = 1.993570.$$

   Suponiendo que $M = N_1(h) + K_1h^2 + K_2h^4 + K_3h^6 + K_4h^8 + O(h^{10})$, construya una tabla de interpolación para determinar $N_4(h)$.
6. Los datos siguientes sirven para aproximar la integral

$$M = \int_0^{3\pi/2} \cos x\, dx.$$

$$N_1(h) = 2.356194, \quad N_1\left(\frac{h}{2}\right) = -0.4879837,$$

$$N_1\left(\frac{h}{4}\right) = -0.8815732, \quad N_1\left(\frac{h}{8}\right) = -0.9709157.$$

   Suponga que existe una fórmula como la que viene en el ejercicio 5 y determine $N_4(h)$.
7. Demuestre que la fórmula de cinco puntos de la ecuación (4.6) aplicada a $f(x) = xe^x$ a $x_0 = 2.0$ da $N_2(0.2)$ en la tabla 4.6 cuando $h = 0.1$ y $N_2(0.1)$ cuando $h = 0.05$.
8. La fórmula de la diferencia progresiva puede expresarse como

$$f'(x_0) = \frac{1}{h}[f(x_0 + h) - f(x_0)] - \frac{h}{2} f''(x_0) - \frac{h^2}{6} f'''(x_0) + O(h^3).$$

   Utilice la extrapolación para derivar una fórmula $O(h^3)$ para $f'(x_0)$.
9. Suponga que $N(h)$ es una aproximación de M para toda $h > 0$ y que

$$M = N(h) + K_1h + K_2h^2 + K_3h^3 + \cdots,$$

para algunas constantes, $K_1, K_2, K_3, \ldots$ Utilice los valores $N(h)$, $N\left(\frac{h}{3}\right)$ y $N\left(\frac{h}{9}\right)$ para producir una aproximación $O(h^3)$ de M.

10. Suponga que $N(h)$ es una aproximación de M para toda $h > 0$ y que

$$M = N(h) + K_1h^2 + K_2h^4 + K_3h^6 + \cdots,$$

para algunas constantes $K_1, K_2, K_3, \ldots$ Utilice los valores $N(h)$, $N\left(\frac{h}{3}\right)$ y $N\left(\frac{h}{9}\right)$ para producir la aproximación $O(h^6)$ de M.

11. En cálculo aprendimos que $e = \lim_{h\to 0}(1 + h)^{1/h}$.

a. Determine las aproximaciones de e correspondientes a $h = 0.04$, 0.02 y 0.01.

b. Use la extrapolación en las aproximaciones, suponiendo que existen las constantes K_1, $K_2, \ldots$, con

$$e = (1 + h)^{1/h} + K_1h + K_2h^2 + K_3h^3 + \cdots,$$

para producir una aproximación $O(h^3)$ de e, donde $h = 0.04$.

c. ¿Piensa que el supuesto de la parte (b) es correcto?

12. a. Demuestre que

$$\lim_{h\to 0}\left(\frac{2+h}{2-h}\right)^{1/h} = e.$$

b. Calcule las aproximaciones de e aplicando la fórmula $N(h) = \left(\frac{2+h}{2-h}\right)^{1/h}$, para $h = 0.04$, 0.02 y 0.01.

c. Suponga que $e = N(h) + K_1h + K_2h^2 + K_3h^3 + \cdots$. Utilice la extrapolación, al menos con 16 dígitos de precisión, para calcular una aproximación $O(h^3)$ de e con $h = 0.04$ ¿Cree que el supuesto es correcto?

d. Demuestre que $N(-h) = N(h)$.

e. Utilice la parte (d) para demostrar que $K_1 = K_3 = K_5 = \cdots = 0$ en la fórmula

$$e = N(h) + K_1h + K_2h^2 + K_3h^3 + K_4h^4 + K_5h^5 + \cdots,$$

de modo que la fórmula se reduce a

$$e = N(h) + K_2h^2 + K_4h^4 + K_6h^6 + \cdots.$$

f. Use los resultados de la parte (e) y una extrapolación para calcular una aproximación $O(h^6)$ de e con $h = 0.04$.

13. Suponga que la tabla de extrapolación siguiente se elaboró para aproximar el número M con $M = N_1(h) + K_1h^2 + K_2h^4 + K_3h^6$:

$N_1(h)$		
$N_1\left(\frac{h}{2}\right)$	$N_2(h)$	
$N_1\left(\frac{h}{4}\right)$	$N_2\left(\frac{h}{2}\right)$	$N_3(h)$

a. Demuestre que el polinomio interpolante lineal $P_{0,1}(h)$ a través de $(h^2, N_1(h))$ y $(h^2/4, N_1(h/2))$ satisface $P_{0,1}(0) = N_2(h)$. De manera semejante demuestre que $P_{1,2}(0) = N_2(h/2)$.

b. Demuestre que el polinomio interpolante lineal $P_{0,2}(h)$ a través de $(h^4, N_2(h))$ y $(h^4/16, N_2(h/2))$ satisface $P_{0,2}(0) = N_3(h)$.

14. Suponga que $N_1(h)$ es una fórmula que produce las aproximaciones $O(h)$ a un número M y que

$$M = N_1(h) + K_1h + K_2h^2 + \cdots,$$

para un conjunto de constantes positivas $K_1, K_2, \ldots$. Entonces $N_1(h), N_1(h/2), N_1(h/4), \ldots$ son todas cotas inferiores de M. ¿Qué puede decirse sobre las aproximaciones extrapoladas $N_2(h)$, $N_3(h), \ldots$?

15. En el año 200 antes de Cristo, Arquímedes utilizó los semiperímetros de polígonos regulares con k lados que inscriben y circunscriben el círculo unitario para aproximar π, es decir, la circunferencia de un semicírculo. Se puede usar la geometría para demostrar que la secuencia de semiparámetros inscritos y circunscritos $\{p_k\}$ y $\{P_k\}$, respectivamente, satisfacen

$$p_k = k \operatorname{sen}\left(\frac{\pi}{k}\right) \quad \text{y} \quad P_k = k \tan\left(\frac{\pi}{k}\right),$$

con $p_k < \pi < P_k$ siempre que $k \geq 4$.

a. Demuestre que $p_4 = 2\sqrt{2}$ y $P_4 = 4$.

b. Demuestre que para $k \geq 4$, las sucesiones satisfacen las relaciones de recurrencia.

$$P_{2k} = \frac{2p_k P_k}{p_k + P_k} \quad \text{y} \quad p_{2k} = \sqrt{p_k P_{2k}}.$$

c. Aproxime π con una exactitud de 10^{-4} calculando para ello p_k y P_k hasta que $P_k - p_k < 10^{-4}$.

d. Utilice la serie de Taylor para demostrar que

$$\pi = p_k + \frac{\pi^3}{3!}\left(\frac{1}{k}\right)^2 - \frac{\pi^5}{5!}\left(\frac{1}{k}\right)^4 + \cdots$$

y

$$\pi = P_k - \frac{\pi^3}{3}\left(\frac{1}{k}\right)^2 + \frac{2\pi^5}{15}\left(\frac{1}{k}\right)^4 - \cdots.$$

e. Utilice la extrapolación con $h = 1/k$ para una mejor aproximación de π.

4.3 Elementos de la integración numérica

A menudo es necesario evaluar la integral definida de una función que no tiene una antiderivada explícita, o cuya antiderivada no es fácil de obtener. El método básico con que se aproxima $\int_a^b f(x)\,dx$ recibe el nombre de **cuadratura numérica** y emplea una suma del tipo

$$\sum_{i=0}^{n} a_i f(x_i)$$

para aproximar $\int_a^b f(x)\,dx$.

El método de la cuadratura que presentamos en esta sección se basa en los polinomios interpolantes descritos en el capítulo 3. Primero seleccionamos un conjunto de nodos dis-

tintos $\{x_0, \ldots, x_n\}$ del intervalo $[a, b]$. Después integramos el polinomio interpolante de Lagrange

$$P_n(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$$

y su término de error de truncamiento en $[a, b]$ para obtener

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \sum_{i=0}^{n} f(x_i)\,L_i(x)dx + \int_a^b \prod_{i=0}^{n}(x - x_i)\,\frac{f^{(n+1)}(\xi(x))}{(n+1)!}\,dx\\ &= \sum_{i=0}^{n} a_i f(x_i) + \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x - x_i) f^{(n+1)}(\xi(x))\,dx,\end{aligned}$$

donde $\xi(x)$ se encuentra en $[a, b]$ para cada x y

$$a_i = \int_a^b L_i(x)\,dx, \quad \text{para cada } i = 0, 1, \ldots, n.$$

Por tanto, la fórmula de la cuadratura es

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

con un error dado por

$$E(f) = \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x - x_i) f^{(n+1)}(\xi(x))\,dx.$$

Antes de explicar la situación general de las fórmulas de cuadratura, estudiaremos las que se obtienen utilizando el primer y segundo polinomios de Lagrange con nodos igualmente espaciados. Esto nos da la **regla del trapecio** y la **regla de Simpson**, que suelen estudiarse en los cursos de cálculo.

Para derivar la regla del trapecio para aproximar $\int_a^b f(x)\,dx$, sean $x_0 = a$, $x_1 = b$, $h = b - a$ y usaremos el polinomio lineal de Lagrange:

$$P_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1).$$

Luego,

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_{x_0}^{x_1} \left[\frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1)\right]dx \qquad (4.21)\\ &\quad + \frac{1}{2}\int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)\,dx.\end{aligned}$$

Dado que $(x - x_0)\,(x - x_1)$ no cambió de signo en $[x_0, x_1]$, podemos aplicar el teorema de valor medio ponderado de las integrales al término de error a fin de obtener, para algún ξ en (x_0, x_1):

$$\int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)\,dx = f''(\xi)\int_{x_0}^{x_1}(x - x_0)(x - x_1)dx$$
$$= f''(\xi)\left[\frac{x^3}{3} - \frac{(x_1 + x_0)}{2}x^2 + x_0x_1x\right]_{x_0}^{x_1}$$
$$= -\frac{h^3}{6}f''(\xi).$$

En consecuencia, la ecuación (4.21) implica que

$$\int_a^b f(x)\,dx = \left[\frac{(x - x_1)^2}{2(x_0 - x_1)}f(x_0) + \frac{(x - x_0)^2}{2(x_1 - x_0)}f(x_1)\right]_{x_0}^{x_1} - \frac{h^3}{12}f''(\xi)$$
$$= \frac{(x_1 - x_0)}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12}f''(\xi).$$

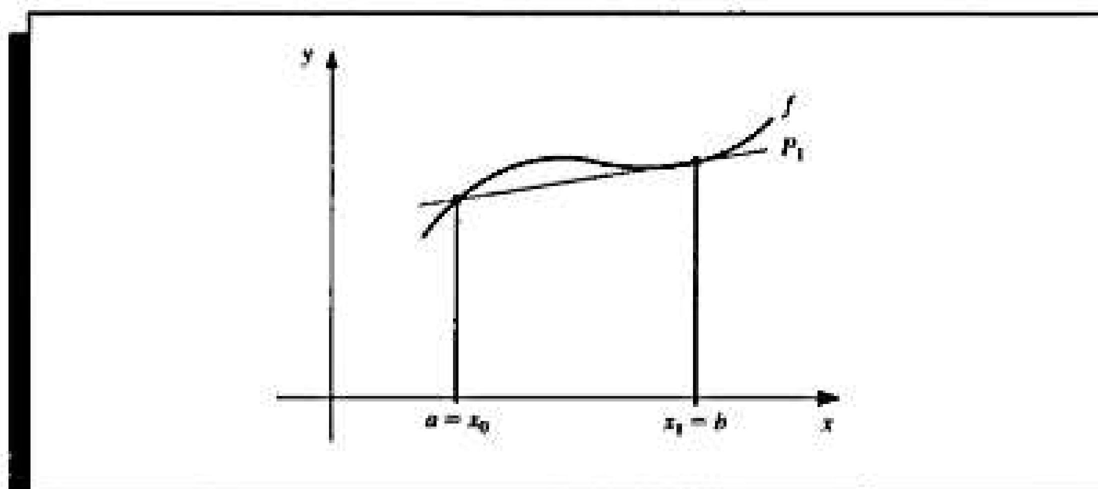
Puesto que $h = x_1 - x_0$, tenemos la siguiente regla:

Regla del trapecio:

$$\int_a^b f(x)\,dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12}f''(\xi).$$

Esta fórmula se llama regla del trapecio porque, cuando f es una función con valores positivos, aproximamos $\int_a^b f(x)\,dx$ por el área de un trapecio, como se muestra en la figura 4.3.

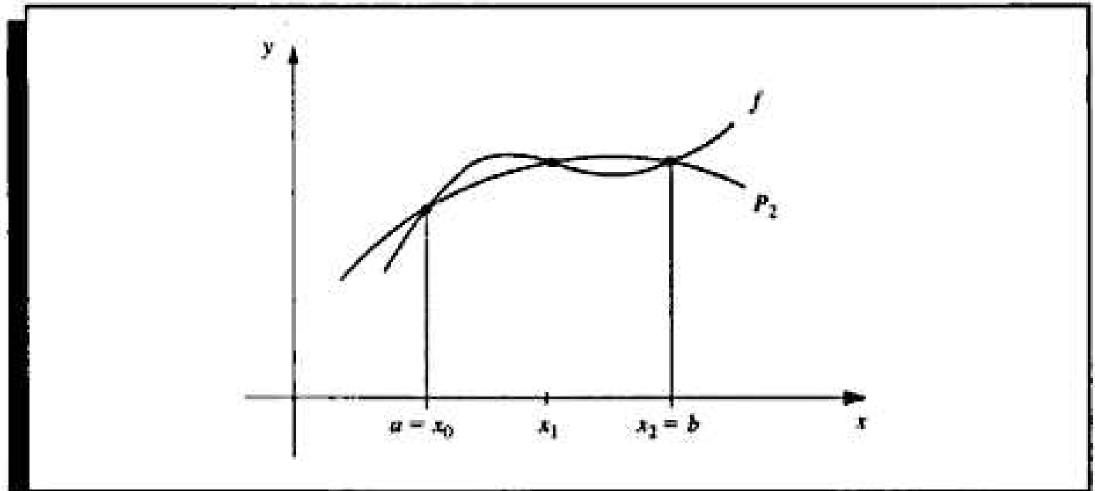
Figura 4.3



Como el término de error de la regla del trapecio contiene f'', la regla da el resultado exacto cuando se aplica a una función cuya segunda derivada sea cero, es decir, cualquier polinomio de grado 1 o menos.

La regla de Simpson se obtiene al integrar en $[a, b]$ el segundo polinomio de Lagrange con los nodos $x_0 = a$, $x_2 = b$ y $x_1 = a + h$, donde $h = (b - a)/2$. (Véase Fig. 4.4.)

Figura 4.4



Por tanto,

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_2}\left[\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1)\right.$$
$$\left. + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2)\right]dx$$
$$+ \int_{x_0}^{x_2}\frac{(x-x_0)(x-x_1)(x-x_2)}{6}f^{(3)}(\xi(x))\,dx.$$

Sin embargo, al deducir la regla de Simpson de esta manera, únicamente se obtiene un término de error $O(h^4)$ que contiene $f^{(3)}$. Si abordamos el problema en otra forma, podemos derivar un término de orden superior que incluya $f^{(4)}$.

Para explicar con un ejemplo esta fórmula alterna, supongamos que f se desarrolla mediante el tercer polinomio de Taylor alrededor de x_1. Entonces, para cada x en $[x_0, x_2]$, existe un número $\xi(x)$ en (x_0, x_2) con

$$f(x) = f(x_1) + f'(x_1)(x-x_1) + \frac{f''(x_1)}{2}(x-x_1)^2 + \frac{f'''(x_1)}{6}(x-x_1)^3 + \frac{f^{(4)}(\xi(x))}{24}(x-x_1)^4$$

y

$$\int_{x_0}^{x_2} f(x)\,dx = \left[f(x_1)(x-x_0) + \frac{f'(x_1)}{2}(x-x_1)^2 + \frac{f''(x_1)}{6}(x-x_1)^3\right. \tag{4.22}$$
$$\left. + \frac{f'''(x_1)}{24}(x-x_1)^4\right]_{x_0}^{x_2} + \frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi(x))(x-x_1)^4\,dx.$$

Puesto que $(x - x_1)^4$ nunca es negativo en $[x_0, x_2]$, el teorema del valor medio ponderado de las integrales implica que

$$\frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi(x))(x-x_1)^4\,dx = \frac{f^{(4)}(\xi_1)}{24}\int_{x_0}^{x_2}(x-x_1)^4\,dx = \frac{f^{(4)}(\xi_1)}{120}(x-x_1)^5\Big]_{x_0}^{x_2},$$

para algún número ξ_1 en (x_0, x_2).

Sin embargo, $h = x_2 - x_1 = x_1 - x_0$, así que

$$(x_2 - x_1)^2 - (x_0 - x_1)^2 = (x_2 - x_1)^4 - (x_0 - x_1)^4 = 0,$$

y en cambio

$$(x_2 - x_1)^3 - (x_0 - x_1)^3 = 2h^3 \quad \text{y} \quad (x_2 - x_1)^5 - (x_0 - x_1)^5 = 2h^5.$$

En consecuencia, podemos reescribir la ecuación (4.22) como

$$\int_{x_0}^{x_2} f(x)dx = 2h\,f(x_1) + \frac{h^3}{3} f''(x_1) + \frac{f^{(4)}(\xi_1)}{60} h^5.$$

Si ahora reemplazamos $f''(x_1)$ por la aproximación de la ecuación (4.9) de la sección 4.1, tendremos

$$\int_{x_0}^{x_2} f(x)\,dx = 2h\,f(x_1) + \frac{h^3}{3}\left\{\frac{1}{h_2}[f(x_0) - 2f(x_1) + f(x_2)] - \frac{h^2}{12} f^{(4)}(\xi_2)\right\} + \frac{f^{(4)}(\xi_1)}{60} h^5$$

$$= \frac{h}{3}\,[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{12}\left[\frac{1}{3} f^{(4)}(\xi_2) - \frac{1}{5} f^{(4)}(\xi_1)\right].$$

Con métodos alternos (consúltese el ejercicio 18) podemos demostrar que en esta expresión los valores ξ_1 y ξ_2 pueden reemplazarse por un valor común ξ en (x_0, x_2). Esto nos da la regla de Simpson.

Regla de Simpson:

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}\,[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90} f^{(4)}(\xi).$$

Dado que el término de error contiene la cuarta derivada de f, la regla de Simpson proporciona resultados exactos al aplicarla a un polinomio cualquiera de grado tres o de grado menor.

EJEMPLO 1 La regla del trapecio para una función f en el intervalo [0, 2] es

$$\int_0^2 f(x)\,dx \approx f(0) + f(2),$$

y la regla de Simpson para f en [0, 2] es

$$\int_0^2 f(x)\,dx \approx \frac{1}{3}\,[f(0) + 4f(1) + f(2)].$$

En la tabla 4.7 se resumen los resultados, con tres decimales, para algunas funciones elementales. Adviértase que en todos los casos la regla de Simpson es mucho mejor. ■

Tabla 4.7

$f(x)$	x^2	x^4	$1/(x+1)$	$\sqrt{1+x^2}$	sen x	e^x
Valores exactos	2.667	6.400	1.099	2.958	1.416	6.389
Trapezoidal	4.000	16.000	1.333	3.326	0.909	8.389
De Simpson	2.667	6.667	1.111	2.964	1.425	6.421

La deducción normal de las fórmulas del error de cuadratura se basa en determinar la clase de polinomios con los cuales estas fórmulas producen resultados exactos. La definición siguiente sirve para facilitar la explicación de esta derivación.

Definición 4.1 El **grado de exactitud** o **precisión** de una fórmula de cuadratura es el entero positivo más grande n, tal que la fórmula sea exacta para x^k, cuando $k = 0, 1, \ldots, n$. ■

La definición 4.1 implica que las reglas del trapecio y de Simpson tienen, respectivamente, un grado de precisión de uno y tres.

La integración y la suma son operaciones lineales, esto es,

$$\int_a^b (\alpha f(x) + \beta g(x))\, dx = \alpha \int_a^b f(x)\, dx + \beta \int_a^b g(x)\, dx$$

y

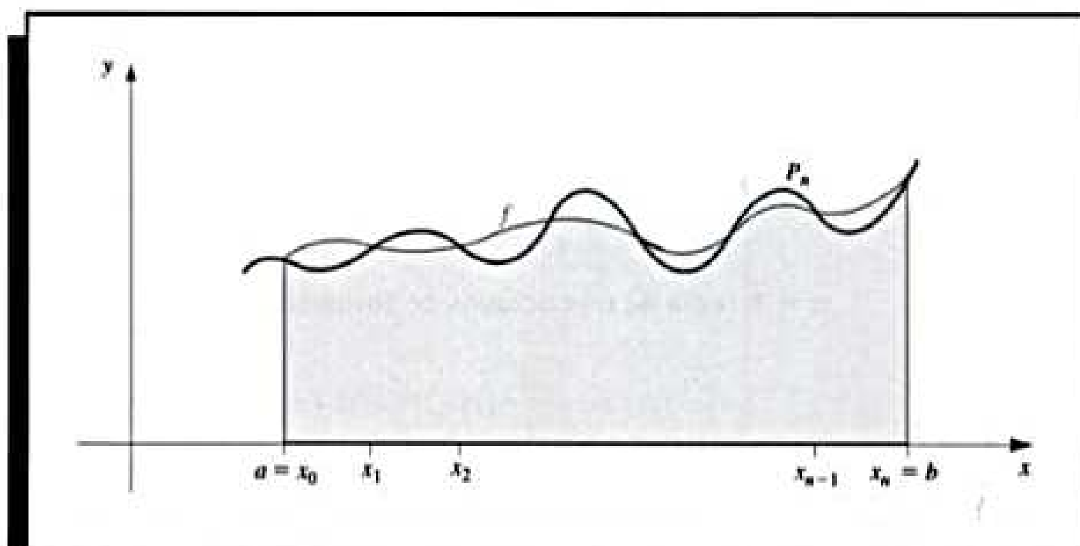
$$\sum_{i=0}^{n} (\alpha f(x_i) + \beta g(x_i)) = \alpha \sum_{i=0}^{n} f(x_i) + \beta \sum_{i=0}^{n} g(x_i),$$

para cada par de funciones integrables f y g, y para cada par de constantes reales α y β. Esto significa (véase el ejercicio 19) que el grado de precisión de una fórmula de cuadratura será n si y sólo si el error $E(P(x)) = 0$ para todos los polinomios $P(x)$ de grado $k = 0, 1, \ldots, n$, pero $E(P(x)) \neq 0$ para algún polinomio $P(x)$ de grado $n + 1$.

Las reglas del trapecio y de Simpson son ejemplos de una clase de métodos denominados fórmulas de Newton-Cotes. Existen dos categorías de fórmulas de Newton-Cotes, abiertas y cerradas.

La *fórmula cerrada de* $(n + 1)$ *puntos de Newton-Cotes* utiliza los nodos $x_i = x_0 + ih$, para $i = 0, 1, \ldots, n$, donde $x_0 = a$, $x_n = b$ y $h = (b - a)/n$. (Véase la figura 4.5.) A esta fórmula se le llama cerrada, porque los extremos del intervalo cerrado $[a, b]$ se incluyen como nodos. La fórmula adopta la forma

Figura 4.5



toma la forma

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

donde

$$a_i = \int_{x_0}^{x_n} L_i(x)\,dx = \int_{x_0}^{x_n} \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}\,dx.$$

En el teorema siguiente se detalla el análisis de error asociado a las fórmulas cerradas de Newton-Cotes. Para una demostración de este teorema consulte [IK, p. 313].

Teorema 4.2 Supongamos que $\sum_{i=0}^{n} a_i f(x_i)$ denota la fórmula cerrada de $(n + 1)$ puntos de Newton-Cotes, con $x_0 = a$, $x_n = b$ y $h = (b - a)/n$. Existe una $\xi \in (a, b)$ para la cual

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_0^n t^2(t-1)\cdots(t-n)\,dt,$$

si n es par, si $f \in C^{n+2}[a, b]$, y si

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_0^n t(t-1)\cdots(t-n)\,dt,$$

si n es impar y si $f \in C^{n+1}[a, b]$. ■

Nótese que, cuando n es un entero par, el grado de precisión es $n + 1$, aunque el polinomio de interpolación es, como máximo, de grado n. En el caso en que n es impar, la segunda parte del teorema indica que el grado de precisión será apenas n.

A continuación se incluyen algunas de las **fórmulas cerradas comunes de Newton-Cotes:**

$n = 1$: regla del trapecio

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12} f''(\xi), \quad \text{donde} \quad x_0 < \xi < x_1. \tag{4.23}$$

$n = 2$: regla de Simpson

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90} f^{(4)}(\xi), \quad \text{donde} \quad x_0 < \xi < x_2. \tag{4.24}$$

$n = 3$: regla de tres octavos de Simpson

$$\int_{x_0}^{x_3} f(x)\,dx = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] - \frac{3h^5}{80} f^{(4)}(\xi), \tag{4.25}$$

donde $x_0 < \xi < x_3$.

n = 4:

$$\int_{x_0}^{x_4} f(x)\,dx = \frac{2h}{45}\left[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)\right] - \frac{8h^7}{945} f^{(6)}(\xi),$$

$$\text{donde} \qquad x_0 < \xi < x_4. \qquad (4.26)$$

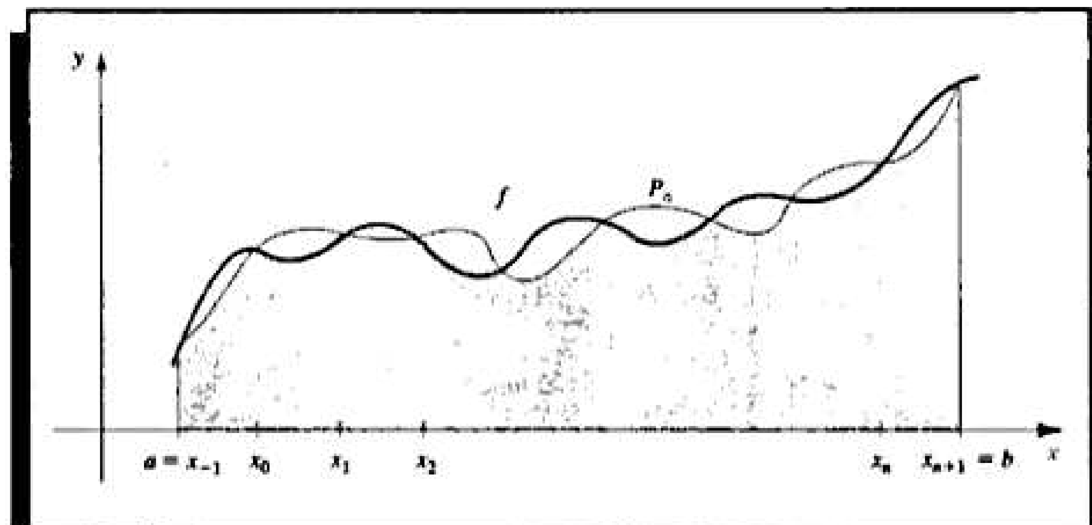
En las *fórmulas abiertas de Newton-Cotes*, los nodos $x_i = x_0 + ih$ se usan para cada $i = 0, 1, \ldots, n$, donde $h = (b - a)/(n + 2)$ y $x_0 = a + h$. Esto implica que $x_n = b - h$, por lo cual marcamos los extremos haciendo $x_{-1} = a$ y $x_{n+1} = b$, como se muestra en la figura 4.6. Las fórmulas abiertas contienen todos los nodos usados para hacer las aproximaciones dentro del intervalo abierto (a, b). Las fórmulas se convierten en

$$\int_a^b f(x)\,dx = \int_{x_{-1}}^{x_{n+1}} f(x)\,dx \approx \sum_{i=0}^{n} a_i f(x_i),$$

donde una vez más

$$a_i = \int_a^b L_i(x)\,dx.$$

Figura 4.6



El teorema siguiente es análogo al 4.2; su demostración se incluye en [IK, p. 314].

Teorema 4.3 Supongamos que $\sum_{i=0}^{n} a_i f(x_i)$ denota la fórmula abierta de $(n + 1)$ puntos de Newton-Cotes, con $x_{-1} = a$, $x_{n+1} = b$, y $h = (b - a)/(n + 2)$. Por tanto, existe $\xi \in (a, b)$ para la cual

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_{-1}^{n+1} t^2 (t-1) \cdots (t-n)\,dt,$$

si n es par y si $f \in C^{n+2}[a, b]$, y si

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_{-1}^{n+1} t(t-1)\cdots(t-n)\,dt,$$

si n es impar y si $f \in C^{n+1}[a, b]$. ■

Algunas de las fórmulas **abiertas de Newton-Cotes** comunes, con sus términos de error son:

$n = 0$: regla del punto medio

$$\int_{x_{-1}}^{x_1} f(x)\,dx = 2h f(x_0) + \frac{h^3}{3} f''(\xi), \quad \text{donde} \quad x_{-1} < \xi < x_1. \tag{4.27}$$

$n = 1$:

$$\int_{x_{-1}}^{x_2} f(x)\,dx = \frac{3h}{2} [f(x_0) + f(x_1)] + \frac{3h^3}{4} f''(\xi), \quad \text{donde} \quad x_{-1} < \xi < x_2. \tag{4.28}$$

$n = 2$:

$$\int_{x_{-1}}^{x_3} f(x)\,dx = \frac{4h}{3} [2f(x_0) - f(x_1) + 2f(x_2)] + \frac{14h^5}{45} f^{(4)}(\xi), \tag{4.29}$$

$$\text{donde} \quad x_{-1} < \xi < x_3.$$

$n = 3$:

$$\int_{x_{-1}}^{x_4} f(x)\,dx = \frac{5h}{24} [11f(x_0) + f(x_1) + f(x_2) + 11f(x_3)] + \frac{95}{144} h^5 f^{(4)}(\xi), \tag{4.30}$$

$$\text{donde} \quad x_{-1} < \xi < x_4.$$

EJEMPLO 2 Al utilizar las fórmulas abiertas y cerradas de Newton-Cotes identificadas como (4.23)−(4.26) y (4.27)−(4.30) para aproximar $\int_0^{\pi/4} \operatorname{sen} x\,dx = 1 - \sqrt{2}/2 \approx 0.29289322$ se obtienen los resultados de la tabla 4.8. ■

Tabla 4.8

n	0	1	2	3	4
Fórmulas cerradas		0.27768018	0.29293264	0.29291070	0.29289318
Error		0.01521303	0.00003942	0.00001748	0.0000004
Fórmulas abiertas	0.30055887	0.29798754	0.29285866	0.29286923	
Error	0.00766565	0.00509432	0.00003456	0.00002399	

CONJUNTO DE EJERCICIOS 4.3

1. Aproxime las siguientes integrales aplicando la regla del trapecio.

   a. $\int_{0.5}^{1} x^4\,dx$
   b. $\int_{0}^{0.5} \frac{2}{x-4}\,dx$
   c. $\int_{1}^{1.5} x^2 \ln x\,dx$
   d. $\int_{0}^{1} x^2 e^{-x}\,dx$
   e. $\int_{1}^{1.6} \frac{2x}{x^2-4}\,dx$
   f. $\int_{0}^{0.35} \frac{2}{x^2-4}\,dx$
   g. $\int_{0}^{\pi/4} x \operatorname{sen} x\,dx$
   h. $\int_{0}^{\pi/4} e^{3x} \operatorname{sen} 2x\,dx$

2. Obtenga una cota del error en el ejercicio 1 aplicando la fórmula de error y compárela con el error real.
3. Repita el ejercicio 1 aplicando la regla de Simpson.
4. Repita el ejercicio 2 usando la regla de Simpson y los resultados del ejercicio 3.
5. Repita el ejercicio 1 aplicando la regla del punto medio.
6. Repita el ejercicio 2 usando la regla del punto medio y los resultados del ejercicio 5.
7. La regla del trapecio aplicada a $\int_0^2 f(x)\,dx$ nos da el valor 4 y la regla de Simpson nos da el valor 2. ¿Qué es $f(1)$?
8. La regla del trapecio aplicada a $\int_0^2 f(x)\,dx$ nos da el valor 5 y la regla del punto medio nos da el valor 4. ¿Qué valor nos da la regla de Simpson?
9. Obtenga el grado de precisión de la fórmula de la cuadratura

$$\int_{-1}^{1} f(x)\,dx = f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right).$$

10. Sean $h = (b-a)/3$, $x_0 = a$, $x_1 = a + h$ y $x_2 = b$. Obtenga el grado de precisión de la fórmula de la cuadratura

$$\int_a^b f(x)\,dx = \frac{9}{4}hf(x_1) + \frac{3}{4}hf(x_2).$$

11. La fórmula de la cuadratura $\int_{-1}^{1} f(x)\,dx = c_0 f(-1) + c_1 f(0) + c_2 f(1)$ es exacta para todos los polinomios de un grado menor o igual a 2. Determine c_0, c_1 y c_2.
12. La fórmula de la cuadratura $\int_0^2 f(x)\,dx = c_0 f(0) + c_1 f(1) + c_2 f(2)$ es exacta para todos los polinomios de un grado menor o igual a 2. Determine c_0, c_1 y c_2.
13. Encuentre las constantes c_0, c_1 y x_1 de modo que la fórmula de la cuadratura

$$\int_0^1 f(x)\,dx = c_0 f(0) + c_1 f(x_1)$$

tenga el grado de precisión más alto posible.

14. Encuentre las constantes x_0, x_1 y c_1 de modo que la fórmula de la cuadratura

$$\int_0^1 f(x)\,dx = \frac{1}{2} f(x_0) + c_1 f(x_1)$$

tenga el grado de precisión más alto posible.

15. Aproxime las siguientes integrales mediante las fórmulas (4.23) a (4.30) ¿Es compatible la exactitud de las aproximaciones con las fórmulas de error? ¿Qué partes de (d) y de (e) dan la mejor aproximación?

a. $\int_0^{0.1} \sqrt{1+x}\, dx$

b. $\int_0^{\pi/2} (\operatorname{sen} x)^2\, dx$

c. $\int_{1.1}^{1.5} e^x\, dx$

d. $\int_1^{10} \frac{1}{x}\, dx$

e. $\int_1^{5.5} \frac{1}{x}\, dx + \int_{5.5}^{10} \frac{1}{x}\, dx$

f. $\int_0^1 x^{1/3}\, dx$

16. Dada la función f en los siguientes valores

x	1.8	2.0	2.2	2.4	2.6
$f(x)$	3.12014	4.42569	6.04241	8.03014	10.46675

aproxime $\int_{1.8}^{2.6} f(x)\, dx$ usando todas las fórmulas de cuadratura incluidas en esta sección que puedan aplicarse.

17. Suponga que los datos del ejercicio 16 tienen errores de redondeo contenidos en la tabla siguiente:

x	1.8	2.0	2.2	2.4	2.6
Error en $f(x)$	2×10^{-6}	-2×10^{-6}	-0.9×10^{-6}	-0.9×10^{-6}	2×10^{-6}

Calcule los errores de redondeo del ejercicio 16.

18. Deduzca la regla de Simpson con el término de error por medio de

$$\int_{x_0}^{x_2} f(x)\, dx = a_0 f(x_0) + a_1 f(x_1) + a_2 f(x_2) + k f^{(4)}(\xi).$$

Obtenga a_0, a_1 y a_2 tomando como base el hecho de que la regla de Simpson es exacta para $f(x) = x^n$ cuando $n = 1$, 2 y 3. Después obtenga k aplicando la fórmula de integración con $f(x) = x^4$.

19. Demuestre el enunciado posterior a la definición 4.1; es decir, demuestre que una fórmula de cuadratura tiene un grado de precisión n, y sólo si el error $E(P(x)) = 0$ para todos los polinomios $P(x)$ de grado $k = 0, 1, \ldots, n$, pero $E(P(x)) \neq 0$ para algún polinomio $P(x)$ de grado $n + 1$.

20. Mediante el teorema 4.2, obtenga la regla de tres octavos de Simpson, ecuación (4.25), con el término de error.

21. Utilice el teorema 4.3 para deducir la ecuación (4.28) con un término de error.

4.4 Integración numérica compuesta

En términos generales, las fórmulas de Newton-Cotes no son adecuadas para utilizarse en intervalos de integración grandes. Para estos casos se requieren fórmulas de grado superior y los valores de sus coeficientes son difíciles de obtener. Además, las fórmulas de Newton-Cotes se basan en los polinomios interpolantes que emplean nodos con espacios iguales, procedimiento que resulta inexacto en intervalos grandes a causa de la naturaleza oscilatoria de los polinomios de grado superior. En esta sección estudiaremos un método *fragmentario* para realizar la integración numérica, en el cual se aplican las fórmulas de Newton-Cotes de bajo orden. Estos son los métodos de mayor uso.

Considere el problema de obtener una aproximación a $\int_0^4 e^x\,dx$. La regla de Simpson con $h = 2$ nos da

$$\int_0^4 e^x\,dx \approx \frac{2}{3}(e^0 + 4e^2 + e^4) = 56.76958.$$

Dado que en este caso la respuesta exacta es $e^4 - e^0 = 53.59815$, el error -3.17143 es mucho mayor del que normalmente aceptaríamos.

Si queremos aplicar un método fragmentario a este problema, dividimos $[0, 4]$ en $[0, 2]$ y en $[2, 4]$ y aplicamos dos veces la regla de Simpson con $h = 1$:

$$\begin{aligned}\int_0^4 e^x\,dx &= \int_0^2 e^x\,dx + \int_2^4 e^x\,dx \\ &\approx \frac{1}{3}\,[e^0 + 4e + e^2] + \frac{1}{3}[e^2 + 4e^3 + e^4] \\ &= \frac{1}{3}\,[e^0 + 4e + 2e^2 + 4e^3 + e^4] \\ &= 53.86385.\end{aligned}$$

El error se redujo a -0.26570. Con estos resultados, subdividimos los intervalos $[0, 2]$ y $[2, 4]$ y aplicamos la regla de Simpson con $h = \frac{1}{2}$, obteniendo así

$$\begin{aligned}\int_0^4 e^x\,dx &= \int_0^1 e^x\,dx + \int_1^2 e^x\,dx + \int_2^3 e^x\,dx + \int_3^4 e^x\,dx \\ &\approx \frac{1}{6}[e_0 + 4e^{1/2} + e] + \frac{1}{6}[e + 4e^{3/2} + e^2] \\ &\quad + \frac{1}{6}[e^2 + 4e^{5/2} + e^3] + \frac{1}{6}[e^3 + 4e^{7/2} + e^4] \\ &= \frac{1}{6}[e^0 + 4e^{1/2} + 2e + 4e^{3/2} + 2e^2 + 4e^{5/2} + 2e^3 + 4e^{7/2} + e^4] \\ &= 53.61622.\end{aligned}$$

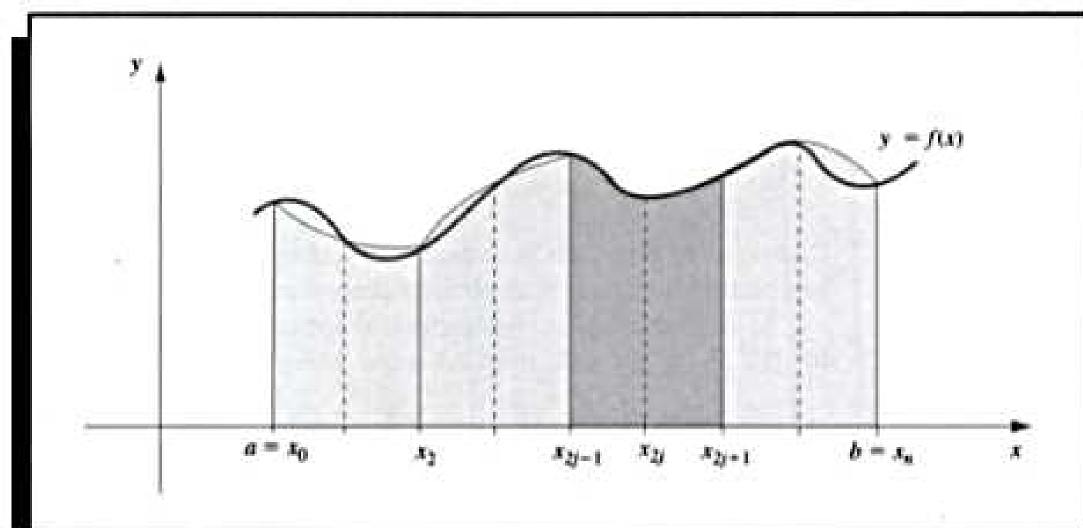
El error de esta aproximación es -0.01807.

Para generalizar este procedimiento, se selecciona un entero par n. Se subdivide el intervalo $[a, b]$ en n subintervalos y se aplica la regla de Simpson en cada par consecutivo de subintervalos. (Véase Fig. 4.7.) Con $h = (b - a)/n$ y $x_j = a + jh$ para cada $j = 0, 1, \ldots, n$, tenemos

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{j=1}^{n/2} \int_{x_{2j-2}}^{x_{2j}} f(x)\,dx \\ &= \sum_{j=1}^{n/2} \left\{ \frac{h}{3}[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})] - \frac{h^5}{90} f^{(4)}(\xi_j) \right\},\end{aligned}$$

para alguna ξ_j con $x_{2j-2} < \xi_j < x_{2j}$, siempre que $f \in C^4[a, b]$. Al aplicar el hecho de que, para cada $j = 1, 2, \ldots, (n/2) - 1$, tenemos que $f(x_{2j})$ aparece en el término correspondiente

Figura 4.7



al intervalo $[x_{2j-2}, x_{2j}]$ y también en el término correspondiente al intervalo $[x_{2j}, x_{2j+2}]$, podemos reducir esta suma a

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(x_0) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(x_n)\right] - \frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j).$$

El error asociado con esta aproximación es

$$E(f) = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j),$$

donde $x_{2j-2} < \xi_j < x_{2j}$ para cada $j = 1, 2, \ldots, n/2$.

Si $f \in C^4[a, b]$, el teorema de valor extremo implica que $f^{(4)}$ asume su máximo y mínimo en $[a, b]$. Puesto que

$$\min_{x\in[a,b]} f^{(4)}(x) \le f^{(4)}(\xi_j) \le \max_{x\in[a,b]} f^{(4)}x,$$

tenemos

$$\frac{n}{2}\min_{x\in[a,b]} f^{(4)}(x) \le \sum_{j=1}^{n/2} f^{(4)}(\xi_j) \le \frac{n}{2}\max_{x\in[a,b]} f^{(4)}(x)$$

y

$$\min_{x\in[a,b]} f^{(4)}(x) \le \frac{2}{n}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) \le \max_{x\in[a,b]} f^{(4)}(x).$$

De acuerdo con el teorema del valor intermedio, existe $\mu \in (a, b)$ tal que

$$f^{(4)}(\mu) = \frac{2}{n}\sum_{j=1}^{n/2} f^{(4)}(\xi_j).$$

Por tanto,

$$E(f) = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) = -\frac{h^5}{180} n f^{(4)}(\mu),$$

o, como $h = (b - a)/n$,

$$E(f) = -\frac{(b-a)}{180} h^4 f^{(4)}(\mu).$$

Las observaciones anteriores producen el siguiente resultado.

Teorema 4.4 Sean $f \in C^4[a, b]$, n par, $h = (b - a)/n$ y $x_j = a + jh$ para cada $j = 0, 1, \ldots, n$. Existe $\mu \in (a, b)$ tal que la **regla compuesta de Simpson** para n subintervalos puede escribirse con su término de error como

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{b-a}{180} h^4 f^{(4)}(\mu).$$ ■

El algoritmo 4.1 usa la regla compuesta de Simpson en n subintervalos. Este es el algoritmo de cuadratura de propósito general que más se usa.

ALGORITMO 4.1

Regla compuesta de Simpson

Para aproximar la integral $I = \int_a^b f(x)\,dx$:

ENTRADA extremos a, b; entero positivo par n.

SALIDA aproximación XI a I.

Paso 1 Tome $h = (b - a)/n$.

Paso 2 Tome $XI0 = f(a) + f(b)$;
$XI1 = 0$; (*Suma de* $f(x_{2i-1})$.)
$XI2 = 0$. (*Suma de* $f(x_{2i})$.)

Paso 3 Para $i = 1, \ldots, n - 1$ efectúe pasos 4 y 5.

Paso 4 Tome $X = a + ih$.

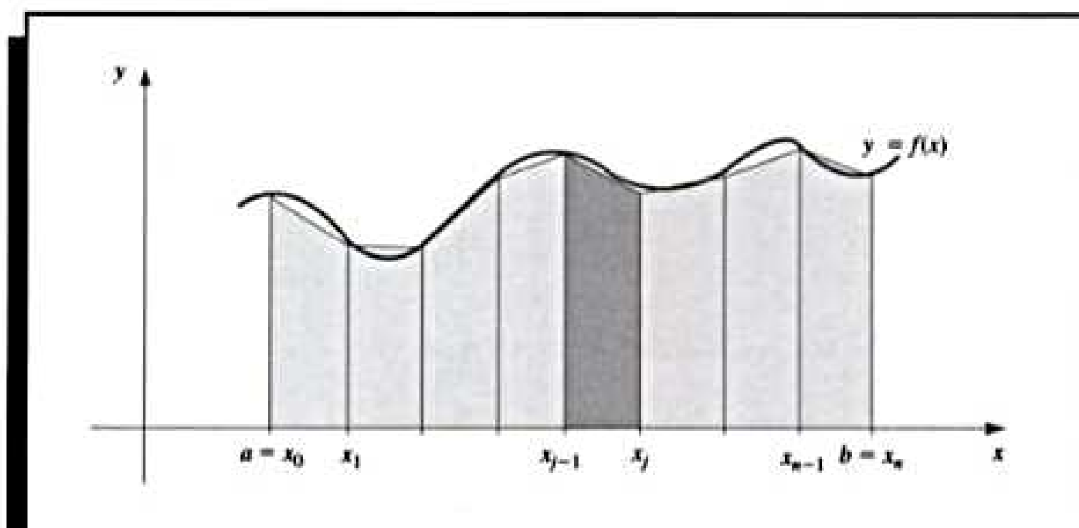
Paso 5 Si i es par entonces tome $XI2 = XI2 + f(X)$
si no, tome $XI1 = XI1 + f(X)$.

Paso 6 Tome $XI = h(XI0 + 2 \cdot XI2 + 4 \cdot XI1)/3$.

Paso 7 SALIDA (XI);
PARAR. ■

El método de subdivisión se puede aplicar a cualquiera de las fórmulas de Newton-Cotes. Las extensiones de las reglas del trapecio (véase Fig. 4.8) y del punto medio se incluyen sin su demostración. La regla del trapecio requiere sólo un intervalo en cada aplicación, por lo cual el entero n puede ser par o impar.

Figura 4.8

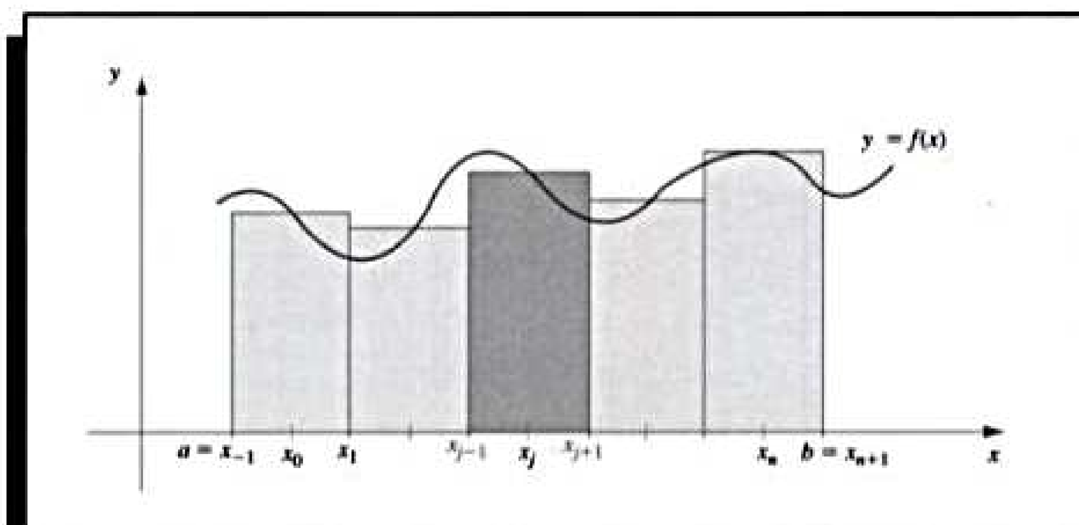


Teorema 4.5 Sean $f \in C^2[a, b]$, $h = (b - a)/n$ y $x_j = a + jh$ para cada $j = 0, 1, \ldots, n$. Existe una $\mu \in (a, b)$ tal que la **regla compuesta del trapecio** para n subintervalos puede escribirse con su término de error como

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \frac{b-a}{12}h^2 f''(\mu). \qquad \blacksquare$$

Para el caso de la regla compuesta del punto medio, nuevamente n debe ser par. (Véase Fig. 4.9.)

Figura 4.9



Teorema 4.6 Sea $f \in C^2[a, b]$, n par, $h = (b - a)/(n + 2)$ y $x_j = a + (j + 1)h$ para cada $j = -1, 0, \ldots, n + 1$. Existe una $\mu \in (a, b)$ tal que la **regla compuesta del punto medio** para $n + 2$ subintervalos puede escribirse con su término de error como sigue

$$\int_a^b f(x)\,dx = 2h\sum_{j=0}^{n/2} f(x_{2j}) + \frac{b-a}{6}h^2 f''(\mu). \qquad \blacksquare$$

EJEMPLO 1 Con la regla compuesta de Simpson, considere el problema de aproximar $\int_0^\pi$ sen $x\,dx$ con un error absoluto menor que 0.00002. Esta regla nos da para algún μ en $(0, \pi)$,

$$\int_0^\pi \operatorname{sen} x\,dx = \frac{h}{3}\left[2\sum_{j=1}^{(n/2)-1} \operatorname{sen} x_{2j} + 4\sum_{j=1}^{n/2} \operatorname{sen} x_{2j-1}\right] - \frac{\pi h^4}{180}\operatorname{sen}\mu.$$

Dado que el error absoluto debe ser menor que 0.00002, la desigualdad

$$\left|\frac{\pi h^4}{180}\operatorname{sen}\mu\right| \le \frac{\pi h^4}{180} = \frac{\pi^5}{180n^4} < 0.00002$$

sirve para determinar n y h. Al completar estos cálculos obtenemos $n \ge 18$. Si $n = 20$, entonces $h = \pi/20$, y la fórmula nos da

$$\int_0^\pi \operatorname{sen} x\,dx \approx \frac{\pi}{60}\left[2\sum_{j=1}^{9} \operatorname{sen}\left(\frac{j\pi}{10}\right) + 4\sum_{j=1}^{10} \operatorname{sen}\left(\frac{(2j-1)\pi}{20}\right)\right] = 2.000006.$$

Para asegurarse del grado de exactitud al usar la regla compuesta del trapecio, se requiere que

$$\left|\frac{\pi h^2}{12}\operatorname{sen}\mu\right| \le \frac{\pi h^2}{12} = \frac{\pi^3}{12n^2} < 0.00002,$$

o que $n \ge 360$. Esto implica realizar un número de cálculos mucho mayor que los que se requieren al aplicar la regla compuesta de Simpson, por lo cual prescindiremos de usar la regla compuesta del trapecio en este problema. Para facilitar la comparación, la regla compuesta del trapecio con $n = 20$ y con $h = \pi/20$ nos da

$$\int_0^\pi \operatorname{sen} x\,dx \approx \frac{\pi}{40}\left[2\sum_{j=1}^{19} \operatorname{sen}\left(\frac{j\pi}{20}\right) + \operatorname{sen} 0 + \operatorname{sen}\pi\right] = \frac{\pi}{40}\left[2\sum_{j=1}^{19} \operatorname{sen}\left(\frac{j\pi}{20}\right)\right]$$
$$= 1.9958860.$$

La respuesta exacta es 2, de manera que la regla de Simpson con $n = 20$ proporciona una respuesta dentro de la cota de error requerida, lo cual evidentemente no sucede en el caso de la regla del trapecio con $n = 20$. ■

La mayor parte de los sistemas de álgebra por computadora incorpora la regla compuesta de Simpson y la del trapecio. En Maple, para tener acceso a la biblioteca donde están definidas, introduzca

```
>with (student);
```

Las llamadas de los métodos son `trapezoid(f,x=a..b,n)` y `Simpson (f,x=a..b,n)`. Para efectos de nuestro ejemplo,

```
>f:=sin(x)
```

$$f := \operatorname{sen}(x)$$

```
>trapezoid(f,x = 0..Pi,20);
```

$$\frac{1}{20}\pi\left(\sum_{i=1}^{19} \operatorname{sen}\left(\frac{1}{20}i\pi\right)\right)$$

```
>evalf(%);
```

$$1.995885974$$

```
>evalf(simpson (f,x = 0..Pi,20));
```

$$2.000006785$$

La regla compuesta del punto medio también aparece en la Biblioteca de Maple y puede utilizarse mediante el comando

```
>evalf(middlesum(f,x=0...Pi,10));
```

que da la aproximación 2.008248408.

Para mostrar el código de Maple correspondiente al método del punto medio, definimos $f(x)$, a, b, n y h con los comandos

```
>f:=x->sin(x);
>a:=0; b:=Pi; n:=18; h:=(b-a)/(n+2);
```

También necesitamos una variable para calcular la suma; inicializamos la variable en 0.

```
>Tot:=0;
```

En Maple, el ciclo controlado por un contador se define como

```
for variable de control del ciclo, from valor inicial to, valor final do
    enunciado;
    enunciado;
    ⋮
    enunciado;
od;
```

j será nuestra variable de control del ciclo, que comienza en 0 y va hasta $n/2 = 9$ en pasos de una unidad. Para cada valor de $j = 0, 1, \ldots, 9$, se recorre el ciclo y se realiza cada cálculo dentro del ciclo hasta encontrar la palabra `od`. Las palabras reservadas implicadas aquí son `for`, `from`, `do` y `od`. Observe que no aparece un punto y coma (;) después de la afirmación `do`.

```
>for j from 0 to n/2 do
>xj:=a+(2*j+1)*h;
>Tot:=evalf(Tot+f(xj))
>od;
```

Esto produce una serie de resultados que culminan en la suma final

$$Tot = \sum_{j=0}^{n/2} f(x_{2j}) = \sum_{j=0}^{9} f(x_{2j}) = 6.392453222.$$

Luego multiplicamos por $2h$ para concluir con el método compuesto del punto medio:

```
>Tot:=evalf(2*h*Tot);
```

$$Tot := 2.008248408$$

Una propiedad importante que comparten todos los métodos de integración compuesta es la estabilidad respecto al error de redondeo. Para demostrarla, suponga que aplicamos la regla compuesta de Simpson con n subintervalos a una función f en $[a, b]$ y que determinamos la cota máxima de dicho error. Supongamos que aproximamos $f(x_i)$ mediante $\tilde{f}(x_i)$ y que

$$f(x_i) = \tilde{f}(x_i) + e_i, \qquad \text{para cada } i = 0, 1, \ldots, n,$$

donde e_i denota el error de redondeo que implica usar $\tilde{f}(x_i)$ para aproximar $f(x_i)$. Entonces, el error acumulado, $e(h)$, en la regla compuesta de Simpson es

$$e(h) = \left|\frac{h}{3}\left[e_0 + 2\sum_{j=1}^{(n/2)-1} e_{2j} + 4\sum_{j=1}^{n/2} e_{2j-1} + e_n\right]\right|$$
$$\leq \frac{h}{3}\left[|e_0| + 2\sum_{j=1}^{(n/2)-1} |e_{2j}| + 4\sum_{j=1}^{n/2} |e_{2j-1}| + |e_n|\right].$$

Si los errores de redondeo están uniformemente acotados por ε, entonces

$$e(h) \leq \frac{h}{3}\left[\varepsilon + 2\left(\frac{n}{2} - 1\right)\varepsilon + 4\left(\frac{n}{2}\right)\varepsilon + \varepsilon\right] = \frac{h}{3}3n\varepsilon = nh\varepsilon.$$

Pero $nh = b - a$, de modo que

$$e(h) \leq (b - a)\varepsilon,$$

es una cota independiente de h (y n). Esto significa que, aunque debamos dividir un intervalo en más partes para garantizar cierta precisión, los cálculos agregados no aumentan el error por redondeo. Este resultado indica que el procedimiento es estable al aproximarse h a cero. Recuerde que no fue así en los procedimientos de diferenciación numérica que estudiamos al inicio del capítulo.

CONJUNTO DE EJERCICIOS 4.4

1. Aplique la regla compuesta del trapecio con los valores indicados de n para aproximar las siguientes integrales.

   a. $\int_1^2 x \ln x\, dx$, $n = 4$
   b. $\int_{-2}^2 x^3 e^x\, dx$, $n = 4$
   c. $\int_0^2 \frac{2}{x^2 + 4}\, dx$, $n = 6$
   d. $\int_0^\pi x^2 \cos x\, dx$, $n = 6$
   e. $\int_0^2 e^{2x} \operatorname{sen} 3x\, dx$, $n = 8$
   f. $\int_1^3 \frac{x}{x^2 + 4}\, dx$, $n = 8$
   g. $\int_3^5 \frac{1}{\sqrt{x^2 - 4}}\, dx$, $n = 8$
   h. $\int_0^{3\pi/8} \tan x\, dx$, $n = 8$

2. Aplique la regla compuesta de Simpson para aproximar las integrales del ejercicio 1.
3. Aplique la regla compuesta del punto medio con $n + 2$ subintervalos para aproximar las integrales del ejercicio 1.

4. Aproxime $\int_0^2 x^2 e^{-x^2}\,dx$ por medio de $h = 0.25$.
   a. Aplique la regla compuesta del trapecio.
   b. Aplique la regla compuesta de Simpson.
   c. Aplique la regla compuesta del punto medio.
5. Suponga que $f(0) = 1, f(0.5) = 2.5, f(1) = 2$ y $f(0.25) = f(0.75) = \alpha$. Determine α si la regla compuesta del trapecio con $n = 4$ da el valor 1.75 para $\int_0^1 f(x)dx$.
6. La regla del punto medio con que se aproxima $\int_{-1}^1 f(x)\,dx$ da el valor 12, la regla compuesta del punto medio con $n = 2$ da 5 y la regla compuesta de Simpson da 6. Aplique el hecho de que $f(-1) = f(1)$ y $f(-0.5) = f(0.5) - 1$ para determinar $f(-1), f(-0.5), f(0), f(0.5)$, y $f(1)$.
7. Determine los valores de n y h que se requieren para aproximar

$$\int_0^2 e^{2x}\,\text{sen}\,3x\,dx$$

con una exactitud de 10^{-4}.
   a. Aplique la regla compuesta del trapecio.
   b. Aplique la regla compuesta de Simpson.
   c. Aplique la regla compuesta del punto medio.
8. Repita el ejercicio 7 con la integral $\int_0^{\pi} x^2 \cos x\,dx$.
9. Determine los valores de n y h que se requieren para aproximar

$$\int_0^2 \frac{1}{x+4}\,dx$$

con una exactitud de 10^{-5} y calcule la aproximación.
   a. Aplique la regla compuesta del trapecio.
   b. Aplique la regla compuesta de Simpson.
   c. Aplique la regla compuesta del punto medio.
10. Repita el ejercicio 9 con la integral $\int_1^2 x \ln x\,dx$.
11. Suponga que f está definida por

$$f(x) = \begin{cases} x^3 + 1, & 0 \le x \le 0.1, \\ 1.001 + 0.03(x - 0.1) + 0.3(x - 0.1)^2 + 2(x - 0.1)^3, & 0.1 \le x \le 0.2, \\ 1.009 + 0.15(x - 0.2) + 0.9(x - 0.2)^2 + 2(x - 0.2)^3, & 0.2 \le x \le 0.3. \end{cases}$$

   a. Investigue la continuidad de las derivadas de f.
   b. Aplique la regla compuesta del trapecio con $n = 6$ para aproximar $\int_0^{0.3} f(x)\,dx$, y estime el error por medio de la cota de error.
   c. Aplique la regla compuesta de Simpson con $n = 6$ para aproximar $\int_0^{0.3} f(x)\,dx$. ¿Son los resultados más exactos que en la parte (b)?
12. Demuestre que el error $E(f)$ de la regla compuesta de Simpson puede aproximarse por medio de

$$-\frac{h^4}{180}\,[f'''(b) - f'''(a)].$$

[*Sugerencia*: $\sum_{j=1}^{n/2} f^{(4)}(\xi_j)\,(2h)$ es una suma de Riemann para $\int_a^b f^{(4)}(x)\,dx$.]

13. a. Con el método usado en el ejercicio 12, derive una estimación para $E(f)$ en la regla compuesta del trapecio.

    b. Repita la parte (a) con la regla compuesta del punto medio.

14. Use las estimaciones de error de los ejercicios 12 y 13 para estimar los errores del ejercicio 8.

15. Use las estimaciones de error de los ejercicios 12 y 13 para estimar los errores del ejercicio 10.

16. En los cursos de cálculo de varias variables y de estadística se demuestra que

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)(x/\sigma)^2}\, dx = 1,$$

para cualquier σ positiva. La función

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)(x/\sigma)^2}$$

es la *función de densidad normal* con la *media* $\mu = 0$ y la *desviación estándar* σ. La probabilidad de que un valor aleatoriamente seleccionado descrito por esta distribución se encuentre en $[a, b]$ está dada por $\int_a^b f(x)\, dx$. Con una exactitud de 10^{-5} aproxime la probabilidad de que un valor aleatoriamente seleccionado descrito por esta distribución se encuentre en

a. $[-\sigma, \sigma]$ b. $[-2\sigma, 2\sigma]$ c. $[-3\sigma, 3\sigma]$

17. Determine con una exactitud de 10^{-6}, la longitud de la gráfica de la elipse que sigue la ecuación $4x^2 + 9y^2 = 36$.

18. Un automóvil recorre una pista de carreras en 84 segundos. Su velocidad en cada intervalo de 6 segundos se determina mediante una pistola de radar y está dada, en pies/s, desde el principio del recorrido, por los datos de la tabla siguiente:

Tiempo	0	6	12	18	24	30	36	42	48	54	60	66	72	78	84
Velocidad	124	134	148	156	147	133	121	109	99	85	78	89	104	116	123

¿Qué longitud tiene la pista?

19. Una partícula de masa m que se desplaza por un fluido está sujeta a una resistencia viscosa R, la cual es una función de la velocidad v. La relación entre la resistencia R, la velocidad v y el tiempo t está dada por la ecuación

$$t = \int_{v(t_0)}^{v(t)} \frac{m}{R(u)}\, du.$$

Suponga que $R(v) = -v\sqrt{v}$ para determinado fluido, donde R se da en newtons y v se da en metros/segundo. Si $m = 10$ kg y si $v(0) = 10$ m/s, aproxime el tiempo que la partícula tarda en reducir su velocidad a $v = 5$ m/s.

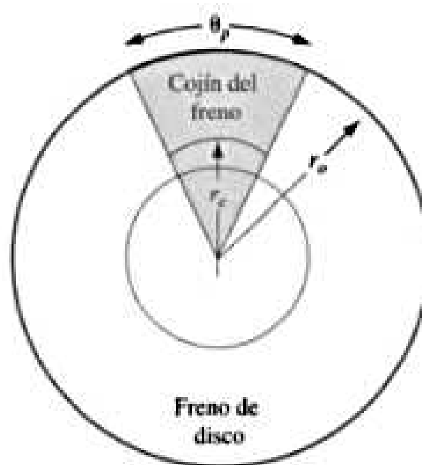
20. Para simular las características térmicas de los frenos de disco (véase la figura anexa), D. A. Secrist y R. W. Hornbeck [SH] tuvieron que aproximar numéricamente la "temperatura exterior promediada del área", T, en el cojín del freno, basándose para ello en la ecuación

$$T = \frac{\int_{r_e}^{r_0} T(r) r\theta_p\, dr}{\int_{r_e}^{r_0} r\theta_p\, dr},$$

donde r_e representa el radio donde comienza el contacto entre cojín y disco, r_0 representa el radio exterior de dicho contacto, θ_p representa el ángulo subtendido por los cojines del freno del sector y $T(r)$ es la temperatura en cada punto del cojín, la cual se obtuvo numéricamente al ana-

lizar la ecuación del calor (véase la sección 12.2). Si $r_e = 0.308$ pies, $r_0 = 0.478$ pies, $\theta_p = 0.7051$ radiantes y si las temperaturas dadas en la tabla siguiente se calcularon en varios puntos del disco, obtenga una aproximación de T.

r (pies)	$T(r)$ (°F)	r (pies)	$T(r)$ (°F)	r (pies)	$T(r)$ (°F)
0.308	640	0.376	1034	0.444	1204
0.325	794	0.393	1064	0.461	1222
0.342	885	0.410	1114	0.478	1239
0.359	943	0.427	1152		



21. Con una exactitud de 10^{-4}, obtenga una aproximación del valor de la integral que se incluye en la aplicación con que inicia este capítulo.

$$\int_0^{48} \sqrt{1 + (\cos x)^2}\, dx.$$

22. La ecuación

$$\int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = 0.45$$

puede resolverse para x aplicando el método de Newton con

$$f(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt - 0.45$$

y con

$$f'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Si queremos evaluar f en la aproximación p_k, necesitamos una fórmula de la cuadratura para aproximar

$$\int_0^{p_k} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

a. Obtenga una solución de $f(x) = 0$ con una exactitud de 10^{-5} aplicando el método de Newton con $p_0 = 0.5$ y la regla compuesta de Simpson.

b. Repita (a) aplicando la regla compuesta del trapecio en vez de la regla compuesta de Simpson.

4.5 Integración de Romberg

En la integración de Romberg se usa la regla compuesta del trapecio para obtener aproximaciones preliminares, y luego el proceso de extrapolación de Richardson para mejorar las aproximaciones. En la sección 4.2 dijimos que la extrapolación de Richardson puede efectuarse en cualquier procedimiento de aproximación de la forma

$$M - N(h) = K_1 h + K_2 h^2 + \cdots + K_n h^n,$$

donde $K_1, K_2, \ldots, K_n$ son constantes y $N(h)$ es una aproximación al valor desconocido M. En esta fórmula el error de truncamiento está dominado por $K_1 h$ cuando h es pequeño y, por tanto, esta fórmula da $O(h)$ aproximaciones. En la extrapolación de Richardson se utiliza una técnica de prorrateo para producir fórmulas con un error de truncamiento de orden superior. En la sección 4.2 vimos cómo esto nos puede servir para obtener aproximaciones de la derivada. En esta sección usaremos la extrapolación para aproximar integrales definidas.

Para comenzar a explicar el método de integración de Romberg, recordemos lo siguiente: la regla compuesta del trapecio para aproximar la integral de una función f en un intervalo $[a, b]$ por medio de m subintervalos es

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + f(b) + 2\sum_{j=1}^{m-1} f(x_j)\right] - \frac{(b-a)}{12}h^2 f''(\mu),$$

donde $a < \mu < b$, $h = (b-a)/m$ y $x_j = a + jh$ para cada $j = 0, 1, \ldots, m$.

Primero obtenemos las aproximaciones mediante la regla compuesta del trapecio, con $m_1, = 1, m_2 = 2, m_3 = 4, \ldots,$ y $m_n = 2^{n-1}$, donde n es un entero positivo. Los valores del tamaño del paso h_k correspondientes a m_k son $h_k = (b-a)/m_k = (b-a)/2^{k-1}$. Con esta notación, la regla del trapecio se expresa como:s

$$\int_a^b f(x)\,dx = \frac{h_k}{2}\left[f(a) + f(b) + 2\left(\sum_{i=1}^{2^{k-1}-1} f(a + ih_k)\right)\right] - \frac{(b-a)}{12}h_k^2 f''(\mu_k). \tag{4.31}$$

donde μ_k es un número en (a, b).

Si introducimos la notación $R_{k,1}$ para denotar la parte de la ecuación (4.31) con que se realiza la aproximación por trapecios, tenemos que

$$R_{1,1} = \frac{h_1}{2}[f(a) + f(b)] = \frac{(b-a)}{2}[f(a) + f(b)];$$

$$R_{2,1} = \frac{h_2}{2}[f(a) + f(b) + 2f(a + h_2)]$$

$$= \frac{(b-a)}{4}\left[f(a) + f(b) + 2f\left(a + \frac{(b-a)}{2}\right)\right]$$

$$= \frac{1}{2}[R_{1,1} + h_1 f(a + h_2)];$$

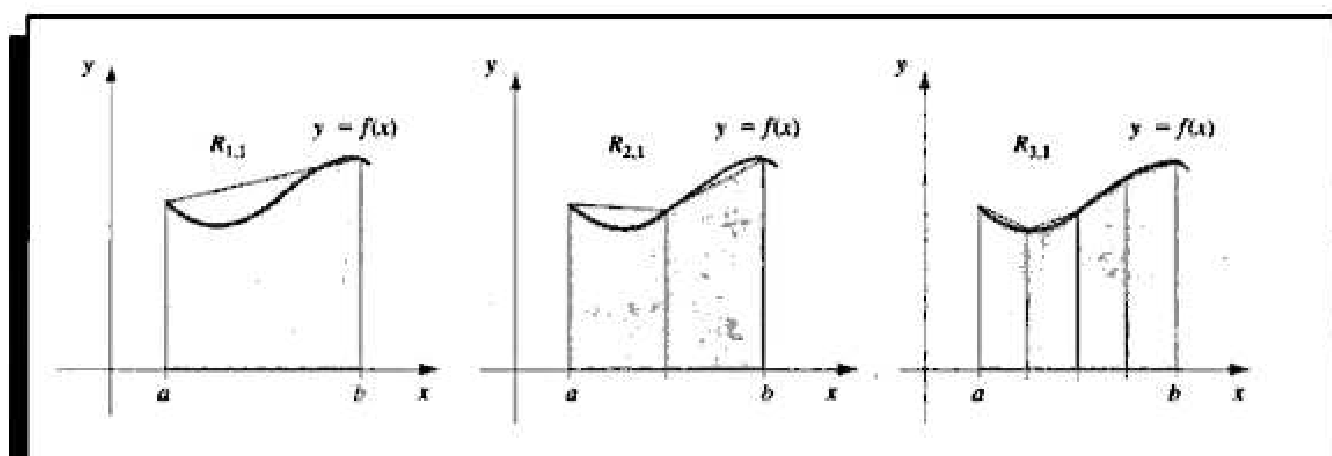
$$R_{3,1} = \frac{1}{2}\{R_{2,1} + h_2[f(a + h_3) + f(a + 3h_3)]\};$$

y, en general (véase Fig. 4.10),

$$R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (2i-1)h_k)\right], \tag{4.32}$$

para cada $k = 2, 3, \ldots, n$. (Véanse los ejercicios 12 y 13).

Figura 4.10



EJEMPLO 1 Al usar la ecuación (4.32) para efectuar el primer paso del método de integración de Romberg para aproximar $\int_0^\pi \operatorname{sen} x\, dx$ con $n = 6$ obtenemos

$$R_{1,1} = \frac{\pi}{2}[\operatorname{sen} 0 + \operatorname{sen} \pi] = 0;$$

$$R_{2,1} = \frac{1}{2}\left[R_{1,1} + \pi \operatorname{sen}\frac{\pi}{2}\right] = 1.57079633;$$

$$R_{3,1} = \frac{1}{2}\left[R_{2,1} + \frac{\pi}{2}\left(\operatorname{sen}\frac{\pi}{4} + \operatorname{sen}\frac{3\pi}{4}\right)\right] = 1.89611890;$$

$$R_{4,1} = \frac{1}{2}\left[R_{3,1} + \frac{\pi}{4}\left(\operatorname{sen}\frac{\pi}{8} + \operatorname{sen}\frac{3\pi}{8} + \operatorname{sen}\frac{5\pi}{8} + \operatorname{sen}\frac{7\pi}{8}\right)\right] = 1.97423160;$$

$R_{5,1} = 1.99357034,$ y $R_{6,1} = 1.99839336.$ ■

El valor correcto de la integral del ejemplo 1 es 2; por tanto, parece que la convergencia es lenta. La extrapolación de Richardson servirá para agilizar la convergencia.

Podemos demostrar, aunque ello no sea fácil (véase [RR, pp. 136-138]) que si $f \in C^\infty[a, b]$, entonces podemos escribir la regla compuesta del trapecio con un término de error alterno en la forma

$$\int_a^b f(x)\,dx - R_{k,1} = \sum_{i=1}^{\infty} K_i h_k^{2i} = K_1 h_k^2 + \sum_{i=2}^{\infty} K_i h_k^{2i}, \tag{4.33}$$

donde K_i para cada i es independiente de h_k sólo en $f^{(2i-1)}(a)$ y $f^{(2i-1)}(b)$.

Con la regla compuesta del trapecio en esta forma, podemos suprimir el término que contiene h_k^2 al combinar esta ecuación con su correspondiente que tiene h_k reemplazada por $h_{k+1} = h_k/2$:

$$\int_a^b f(x)\,dx - R_{k+1,1} = \sum_{i=1}^{\infty} K_i h_{k+1}^{2i} = \sum_{i=1}^{\infty} \frac{K_i h_k^{2i}}{2^{2i}} = \frac{K_1 h_k^2}{4} + \sum_{i=2}^{\infty} \frac{K_i h_k^{2i}}{4^i}. \tag{4.34}$$

Al restar la ecuación (4.33) a cuatro veces la (4.34) y al simplificarla, obtenemos la fórmula $O(h_k^4)$

$$\int_a^b f(x)\,dx - \left[R_{k+1,1} + \frac{R_{k+1,1} - R_{k,1}}{3}\right] = \sum_{i=2}^{\infty} \frac{K_i}{3}\left(\frac{h_k^{2i}}{4^{i-1}} - h_k^{2i}\right)$$

$$= \sum_{i=2}^{\infty} \frac{K_i}{3}\left(\frac{1 - 4^{i-1}}{4^{i-1}}\right) h_k^{2i}.$$

Ahora podemos aplicar la extrapolación a esta fórmula para obtener un resultado $O(h_k^6)$ y así sucesivamente. Para simplificar la notación definimos

$$R_{k,2} = R_{k,1} + \frac{R_{k,1} - R_{k-1,1}}{3},$$

para cada $k = 2, 3, \ldots, n$, y aplicamos el procedimiento de extrapolación de Richardson a estos valores. Continuando esta notación tenemos, para cada $k = 2, 3, 4, \ldots, n$ y $j = 2, \ldots, k$, una fórmula de aproximación $O(h_k^{2j})$ definida por

$$R_{k,j} = R_{k,j-1} + \frac{R_{k,j-1} - R_{k-1,j-1}}{4^{j-1} - 1}. \tag{4.35}$$

Los resultados generados con estas fórmulas se incluyen en la tabla 4.9.

Tabla 4.9

$R_{1,1}$					
$R_{2,1}$	$R_{2,2}$				
$R_{3,1}$	$R_{3,2}$	$R_{3,3}$			
$R_{4,1}$	$R_{4,2}$	$R_{4,3}$	$R_{4,4}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
$R_{n,1}$	$R_{n,2}$	$R_{n,3}$	$R_{n,4}$	$\cdots$	$R_{n,n}$

El método de Romberg tiene la característica adicional de que permite calcular íntegramente un nuevo renglón de la tabla con sólo hacer una aplicación más de la regla compuesta del trapecio, y luego promediar los valores previamente calculados para obtener los

elementos sucesivos del renglón. El método con que se construye una tabla de este tipo calcula los elementos o datos renglón por renglón, es decir, en el orden $R_{1,1}$, $R_{2,1}$, $R_{2,2}$, $R_{3,1}$, $R_{3,2}$, $R_{3,3}$, etc. El algoritmo 4.2 lo describe en forma detallada.

ALGORITMO 4.2

Integración de Romberg

Para aproximar la integral $I = \int_a^b f(x)\,dx$, seleccione un entero $n > 0$.

ENTRADA extremos a, b; entero n.

SALIDA un arreglo R. *(Calcule R por renglones; sólo los 2 últimos renglones se guardan en almacenamiento.)*

Paso 1 Tome $h = b - a$;
$R_{1,1} = \frac{h}{2}(f(a) + f(b))$.

Paso 2 SALIDA ($R_{1,1}$).

Paso 3 Para $i = 2, \ldots, n$ haga pasos 4-8.

Paso 4 Tome $R_{2,1} = \frac{1}{2}\left[R_{1,1} + h\sum_{k=1}^{2^{i-2}} f(a + (k - 0.5)h)\right]$.

(Aproximación con el método del trapecio.)

Paso 5 Para $j = 2, \ldots, i$

set $R_{2,j} = R_{2,j-1} + \frac{R_{2,j-1} - R_{1,j-1}}{4^{j-1} - 1}$. *(Extrapolación.)*

Paso 6 SALIDA ($R_{2,j}$ por $j = 1, 2, \ldots, i$).

Paso 7 Tome $h = h/2$.

Paso 8 Para $j = 1, 2, \ldots, i$ tome $R_{1,j} = R_{2,j}$. *(Actualice el renglón 1 de R.)*

Paso 9 PARAR. ■

EJEMPLO 2 En el ejemplo 1, los valores de $R_{1,1}$ a $R_{6,1}$ se obtuvieron aproximando $\int_0^\pi \operatorname{sen} x\,dx$. En la tabla 4.10, se muestra la tabla de Romberg que resulta al usar el algoritmo 4.2. ■

Aunque la tabla tiene 21 entradas, solo las seis de la primera columna necesitan evaluaciones funcionales, pues éstas son las únicas entradas generadas por la técnica de integración; las demás se obtienen al calcular promedios.

Tabla 4.10

0					
1.57079633	2.09439511				
1.89611890	2.00455976	1.99857073			
1.97423160	2.00026917	1.99998313	2.00000555		
1.99357034	2.00001659	1.99999975	2.00000001	1.99999999	
1.99839336	2.00000103	2.00000000	2.00000000	2.00000000	2.00000000

El algoritmo 4.2 requiere un entero n previamente establecido, para determinar el número de renglones a generar. A menudo conviene más fijar una tolerancia de error de la aproximación y generar n, dentro de una cota superior, hasta que las entradas diagonales consecutivas $R_{n-1,n-1}$ y $R_{n,n}$ concuerden en el margen de tolerancia. Para evitar la posibilidad de que dos elementos de renglón consecutivos concuerden entre sí, pero no con el valor de la integral a aproximar, generamos aproximaciones hasta que no sólo $|R_{n-1,n-1} - R_{n,n}|$ esté dentro de la tolerancia, sino también $|R_{n-2,n-2} - R_{n-1,n-1}|$. Aunque esta medida no es aplicable a todos los casos, nos garantizará que dos conjuntos de aproximaciones generados en forma distinta concuerden dentro del límite de la tolerancia especificada, antes de que aceptemos $R_{n,n}$ como suficientemente exacto.

La integración de Romberg aplicada a f en $[a, b]$, se basa en la suposición de que la regla compuesta del trapecio tiene un término de error que podemos expresar en la forma de la ecuación (4.33); es decir, debemos tener $f \in C^{2k+2}[a, b]$ para poder generar el k-ésimo renglón. Romberg incluye una verificación en cada etapa para cerciorarse de que la suposición se cumple. A estos métodos se les llama *algoritmos cautelosos de Romberg* y se describen en [Joh]. En esa obra se explica también cuando se usa el método de Romberg como procedimiento adaptativo, semejante a la regla adaptativa de Simpson que estudiaremos en la sección 4.6.

CONJUNTO DE EJERCICIOS 4.5

1. Por medio de la integración de Romberg calcule $R_{3,3}$ para las siguientes integrales.

   a. $\int_1^{1.5} x^2 \ln x \, dx$

   b. $\int_0^1 x^2 e^{-x} \, dx$

   c. $\int_0^{0.35} \frac{2}{x^2 - 4} \, dx$

   d. $\int_0^{\pi/4} x^2 \operatorname{sen} x \, dx$

   e. $\int_0^{\pi/4} e^{3x} \operatorname{sen} 2x \, dx$

   f. $\int_1^{1.6} \frac{2x}{x^2 - 4} \, dx$

   g. $\int_3^{3.5} \frac{x}{\sqrt{x^2 - 4}} \, dx$

   h. $\int_0^{\pi/4} (\cos x)^2 \, dx$

2. Calcule $R_{4,4}$ para las integrales del ejercicio 1.

3. Use la integración de Romberg para aproximar las integrales del ejercicio 1 con una exactitud de 10^{-6}. Calcule la tabla de Romberg hasta que $|R_{n-1,n-1} - R_{n,n}| < 10^{-6}$, o hasta que $n = 10$. Compare sus resultados con los valores exactos de las integrales.

4. Aplique la integración de Romberg a las siguientes integrales hasta que $R_{n-1,n-1}$ y $R_{n,n}$ concuerden con una exactitud de 10^{-4}.

   a. $\int_0^1 x^{1/3} \, dx$

   b. $\int_0^{0.3} f(x) \, dx$, donde

$$f(x) = \begin{cases} x^3 + 1, & 0 \le x \le 0.1 \\ 1.001 + 0.03(x - 0.1) + 0.3(x - 0.1)^2 + 2(x - 0.1)^3, & 0.1 < x \le 0.2, \\ 1.009 + 0.15(x - 0.2) + 0.9(x - 0.2)^2 + 2(x - 0.2)^3, & 0.2 < x \le 0.3. \end{cases}$$

5. Use los siguientes datos para aproximar $\int_1^5 f(x)dx$ con la mayor exactitud posible.

x	1	2	3	4	5
$f(x)$	2.4142	2.6734	2.8974	3.0976	3.2804

6. La integración de Romberg sirve para aproximar

$$\int_0^1 \frac{x^2}{1+x^3}dx.$$

Si $R_{11} = 0.250$ y $R_{22} = 0.2315$, ¿qué será R_{21}?

7. La integración de Romberg se usa para aproximar

$$\int_2^3 f(x)dx.$$

Si, $f(2) = 0.51342$, $f(3) = 0.36788$, $R_{31} = 0.43687$, y si $R_{33} = 0.43662$, obtenga $f(2.5)$.

8. La integración de Romberg para aproximar $\int_0^1 f(x)\,dx$ se tiene $R_{11} = 4$ y $R_{22} = 5$. Encuentre $f\left(\frac{1}{2}\right)$

9. La integración de Romberg para aproximar $\int_a^b f(x)\,dx$ se tiene $R_{11} = 8$, $R_{22} = \frac{16}{3}$, y $R_{33} = \frac{208}{45}$. Encuentre R_{31}.

10. Use la integración de Romberg para calcular las siguientes aproximaciones a

$$\int_0^{48} \sqrt{1 + (\cos x)^2}\,dx.$$

[*Nota*: Los resultados de este ejercicio son muy interesantes en caso de que esté usted utilizando un dispositivo que maneje una aritmética entre siete y nueve dígitos.]

a. Determine $R_{1,1}$, $R_{2,1}$, $R_{3,1}$, $R_{4,1}$ y $R_{5,1}$ y utilice estas aproximaciones para predecir el valor de la integral.

b. Determine $R_{2,2}$, $R_{3,3}$, $R_{4,4}$ y $R_{5,5}$, y modifique su predicción.

c. Determine $R_{6,1}$, $R_{6,2}$, $R_{6,3}$, $R_{6,4}$, $R_{6,5}$ y $R_{6,6}$ y modifique su predicción.

d. Determine $R_{7,7}$, $R_{8,8}$, $R_{9,9}$ y $R_{10,10}$ y haga una predicción final.

e. Explique por qué esta integral causa problemas en la integración de Romberg y cómo podemos reformularla para obtener más fácilmente una aproximación exacta.

11. Demuestre que la aproximación obtenida a partir de $R_{k,2}$ es la misma que la dada por la regla compuesta de Simpson que se describe en el teorema 4.4 con $h = h_k$.

12. Demuestre que, para cualquier k,

$$\sum_{i=1}^{2^{k-1}-1} f\left(a + \frac{i}{2}h_{k-1}\right) = \sum_{i=1}^{2^{k-2}} f\left(a + \left(i - \frac{1}{2}\right)h_{k-1}\right) + \sum_{i=1}^{2^{k-2}-1} f(a + ih_{k-1}).$$

13. Use el resultado del ejercicio 12 para verificar la ecuación (4.32); es decir, demuestre que para toda k

$$R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f\left(a + \left(i - \frac{1}{2}\right)h_{k-1}\right)\right].$$

14. En el ejercicio 24 de la sección 1.1 se integró una serie de Maclaurin para aproximar erf(1), donde erf(x) es la función de error de la distribución normal que se define mediante:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

Aproxime erf(1) con una exactitud de 10^{-7}.

4.6 Métodos adaptativos de cuadratura

En las fórmulas compuestas se requiere el uso de nodos equidistantes. Esto no es adecuado cuando se integra una función en un intervalo que contiene regiones con variación funcional muy grande y regiones con variación funcional pequeña. Si el error de distribución va a estar distribuido uniformemente, se requiere un paso de menor tamaño en las regiones de gran variación que en las de menor variación. En este tipo de problema un método eficiente deberá predecir el grado de variación funcional y adaptar el tamaño del paso a las diversas necesidades. Estas técnicas se conocen con el nombre de **métodos adaptativos de cuadratura**. El método que explicamos en este apartado se basa en la regla compuesta de Simpson, pero podemos modificarlo fácilmente para utilizar otros procedimientos compuestos.

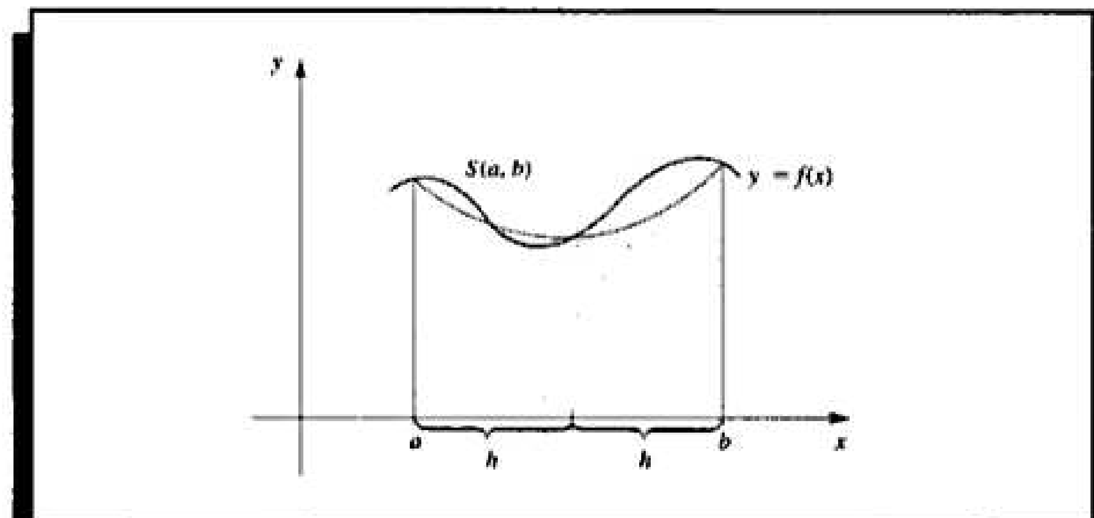
Supóngase que queremos aproximar $\int_a^b f(x)\,dx$ con una tolerancia especificada $\varepsilon>0$. El primer paso del procedimiento consiste en aplicar la regla de Simpson con el tamaño de paso $h(b-a)/2$. Este procedimiento nos da lo siguiente (véase Fig. 4.11):

$$\int_a^b f(x)\,dx = S(a,b) - \frac{h^5}{90} f^{(4)}(\mu), \qquad \text{para algunos } \mu \text{ en } (a,b), \qquad (4.36)$$

donde

$$S(a,b) = \frac{h}{3}\,[f(a) + 4\,f(a+h) + f(b)].$$

Figura 4.11



El siguiente paso consiste en encontrar una forma de estimar la exactitud de nuestra aproximación, en especial una que no requiera determinar $f^{(4)}(\mu)$. Para ello, primero aplicamos la regla compuesta de Simpson al problema con $n = 4$ y el tamaño de paso $(b - a)/4 = h/2$, lo cual nos da

$$\int_a^b f(x)\,dx = \frac{h}{6}\left[f(a) + 4f\left(a + \frac{h}{2}\right) + 2f(a+h) + 4f\left(a + \frac{3h}{2}\right) + f(b)\right]$$

$$-\left(\frac{h}{2}\right)^4 f^{(4)}(\tilde{\mu}), \tag{4.37}$$

para alguna $\tilde{\mu}$ en (a, b). Para simplificar la notación, supongamos que

$$S\left(a, \frac{a+b}{2}\right) = \frac{h}{6}\left[f(a) + 4f\left(a + \frac{h}{2}\right) + f(a+h)\right]$$

y que

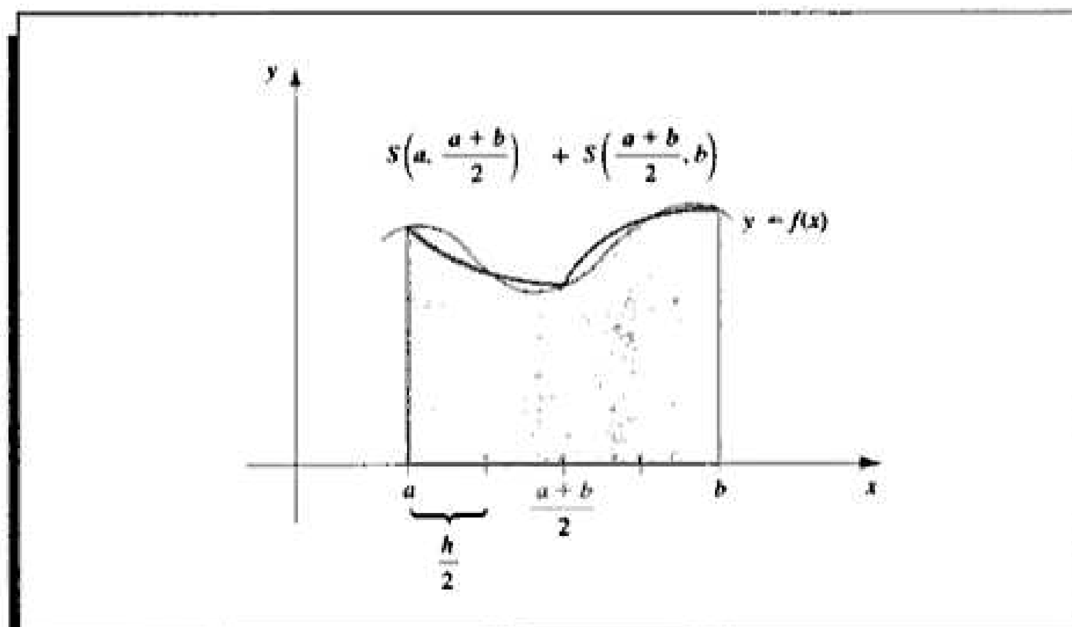
$$S\left(\frac{a+b}{2}, b\right) = \frac{h}{6}\left[f(a+h) + 4f\left(a + \frac{3h}{2}\right) + f(b)\right].$$

Entonces podemos reescribir la ecuación (4.37) (véase la Fig. 4.12) como

$$\int_a^b f(x)\,dx = S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\left(\frac{h^5}{90}\right)f^{(4)}(\tilde{\mu}). \tag{4.38}$$

La estimación de error se obtiene suponiendo que $\mu \approx \tilde{\mu}$ o, más exactamente, que $f^{(4)}(\mu) \approx f^{(4)}(\tilde{\mu})$. El éxito del método depende de la exactitud de esta suposición. Si es exacta, entonces igualar las integrales en las ecuaciones (4.36) y (4.38) implica que

Figura 4.12



$$S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right) - \frac{1}{16}\left(\frac{h^5}{90}\right) f^{(4)}(\tilde{\mu}) \approx S(a, b) - \frac{h^5}{90} f^{(4)}(\mu),$$

así

$$\frac{h^5}{90} f^{(4)}(\mu) \approx \frac{16}{15}\left[S(a, b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right].$$

Al utilizar esta estimación en la ecuación (4.38) obtenemos la estimación de error

$$\left|\int_a^b f(x)\,dx - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right|$$

$$\approx \frac{1}{15}\left|S(a, b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right|.$$

Este resultado significa que $S(a, (a+b)/2) + S((a+b)/2, b)$ aproxima $\int_a^b f(x)\,dx$ unas 15 veces mejor de lo que concuerda con el valor conocido $S(a, b)$. Por tanto, si

$$\left|S(a, b) - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right| < 15\varepsilon, \tag{4.39}$$

esperamos tener

$$\left|\int_a^b f(x)\,dx - S\left(a, \frac{a+b}{2}\right) - S\left(\frac{a+b}{2}, b\right)\right| < \varepsilon, \tag{4.40}$$

y se supone que

$$S\left(a, \frac{a+b}{2}\right) + S\left(\frac{a+b}{2}, b\right)$$

es una aproximación suficientemente exacta a $\int_a^b f(x)\,dx$.

EJEMPLO 1 Para mostrar la exactitud de la estimación del error que se da en las ecuaciones (4.39) y (4.40), consideramos su aplicación a la integral

$$\int_0^{\pi/2} \operatorname{sen} x\,dx = 1.$$

En este caso,

$$S\left(0, \frac{\pi}{2}\right) = \frac{\pi/4}{3}\left[\operatorname{sen} 0 + 4\operatorname{sen}\frac{\pi}{4} + \operatorname{sen}\frac{\pi}{2}\right] = \frac{\pi}{12}(2\sqrt{2} + 1) = 1.002279878$$

y

$$S\left(0, \frac{\pi}{4}\right) + S\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi/8}{3}\left[\operatorname{sen} 0 + 4\operatorname{sen}\frac{\pi}{8} + 2\operatorname{sen}\frac{\pi}{4} + 4\operatorname{sen}\frac{3\pi}{8} + \operatorname{sen}\frac{\pi}{2}\right]$$

$$= 1.000134585,$$

por tanto,

$$\frac{1}{15}\left|S\left(0,\frac{\pi}{2}\right)-S\left(0,\frac{\pi}{4}\right)-S\left(\frac{\pi}{4},\frac{\pi}{2}\right)\right| = 0.000143020.$$

Esto se aproxima muy bien al error real,

$$\left|\int_0^{\pi/2} \operatorname{sen} x\,dx - 1.000134585\right| = 0.000134585,$$

aunque $D_x^4 \operatorname{sen} x = \operatorname{sen} x$ varía significativamente en el intervalo $(0, \pi/2)$. ■

Cuando la estimación del error en (4.39) difiere por más de 15ε no es válida, aplicamos la regla de Simpson de manera individual a los subintervalos $[a, (a+b)/2]$ y $[(a+b)/2, b]$. Después utilizamos el procedimiento de estimación del error para determinar si la aproximación a la integral en cada intervalo se encuentra dentro de una tolerancia de $\varepsilon/2$. De ser así, sumamos las aproximaciones para producir una aproximación a $\int_a^b f(x)\,dx$ con una tolerancia de ε.

En caso de que la aproximación en uno de los subintervalos no se encuentre dentro de la tolerancia $\varepsilon/2$, subdividimos ese subintervalo y repetimos el procedimiento en dos subintervalos para determinar si la aproximación en cada subintervalo tiene una exactitud de $\varepsilon/4$. Continuamos este procedimiento de división en mitades hasta que cada parte esté dentro de la tolerancia requerida. Aunque podemos construir problemas en los que nunca se obtendrá esta tolerancia, el procedimiento suele ser eficaz, porque con cada subdivisión por lo general aumenta la exactitud de la aproximación en un factor de 16, aunque se requiere un factor de mayor precisión de sólo 2.

En el algoritmo 4.3 se detalla este procedimiento adaptativo de la cuadratura para la regla de Simpson, pero se presentan algunos problemas técnicos que requieren que la implantación del método difiera un poco de lo expuesto anteriormente. Por ejemplo, en el paso 1 fijamos la tolerancia en 10ε y no en 15ε, como se indica en la desigualdad (4.39). Esta cota la elegimos conservadoramente para compensar el error de la suposición $f^{(4)}(\mu) \approx f^{(4)}(\tilde{\mu})$. En los problemas en que se sabe que $f^{(4)}$ varía mucho, conviene reducir aún más esta cota.

En una subdivisión, el procedimiento que se incluye en el algoritmo aproxima primero la integral en el subintervalo del extremo izquierdo. Para ello es necesario introducir un procedimiento que almacene y llame eficientemente las evaluaciones funcionales calculadas con anterioridad para los nodos de los subintervalos de la mitad derecha. Los pasos 3, 4 y 5 contienen un procedimiento para apilar, con un indicador que lleva un control de los datos necesarios para calcular la aproximación en el subintervalo contiguo y a la derecha del subintervalo sobre el cual se va a generar la aproximación. El método es más fácil de implantar en una computadora, si se usa un lenguaje de programación que permita la recursión.

ALGORITMO 4.3

Cuadratura adaptativa

Para aproximar la integral $I = \int_a^b f(x)\,dx$ con una tolerancia dada:

ENTRADA extremos a, b; tolerancia TOL; límite al número de niveles N.
SALIDA aproximación APP o mensaje que N fue excedido.

Paso 1 Tome $APP = 0$;
$i = 1$;
$TOL_i = 10\,TOL$;

$a_i = a$;
$h_i = (b - a)/2$;
$FA = f(a)$;
$FC_i = f(a + h_i)$;
$FB_i = f(b)$;
$S_i = h_i(FA_i + 4FC_i + FB_i)/3$; *(Aproximación a partir del método de Simpson para el intervalo completo.)*
$L_i = 1$.

Paso 2 Mientras $i > 0$ haga pasos 3-5.

Paso 3 Tome $FD = f(a_i + h_i/2)$;
$FE = f(a_i + 3h_i/2)$;
$S1 = h_i(FA_i + 4FD + FC_i)/6$; *(Aproximaciones a partir del método de Simpson para mitades de subintervalos.)*
$S2 = h_i(FC_i + 4FE + FB_i)/6$;
$v_1 = a_i$; *(Guarde los datos en este nivel.)*
$v_2 = FA_i$;
$v_3 = FC_i$;
$v_4 = FB_i$;
$v_5 = h_i$;
$v_6 = TOL_i$;
$v_7 = S_i$;
$v_8 = L_i$.

Paso 4 Tome $i = i - 1$. *(Elimine el nivel.)*

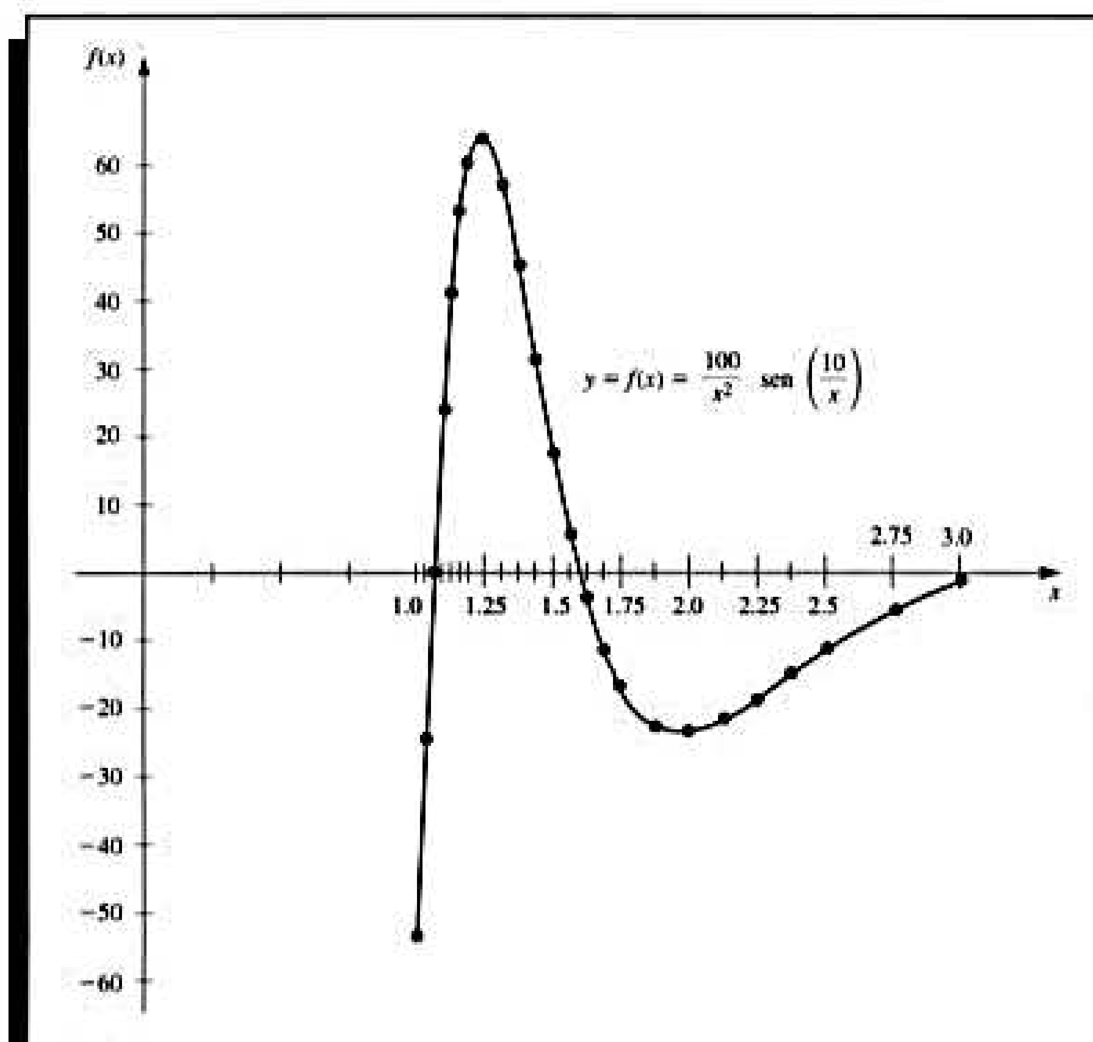
Paso 5 Si $|S1 + S2 - v_7| < v_6$
entonces tome $APP = APP + (S1 + S2)$
si no
si $(v_8 \geq N)$
entonces
SALIDA ('NIVEL EXCEDIDO'); *(El procedimiento falla.)*
PARAR.
si no *(Agregue un nivel.)*
tome $i = i + 1$; *(Datos para la mitad del subintervalo de la derecha.)*
$a_i = v_1 + v_5$;
$FAi = v_3$;
$FC_i = FE$;
$FB_i = v_4$;
$h_i = v_5/2$;
$TOL_i = v_6/2$;
$S_i = S2$;
$L_i = v_8 + 1$;
tome $i = i + 1$; *(Datos para la mitad del subintervalo de la izquierda.)*
$a_i = v_1$;
$FA_i = v_2$;
$FC_i = FD$;
$FB_i = v_3$;
$h_i = h_{i-1}$;

$TOL_i = TOL_i - 1$;
$S_i = S1$;
$L_i = L_{i-1}$.

Paso 6 SALIDA (*APP*); (*APP aproxima a I con precisión TOL.*)
PARAR. ■

EJEMPLO 2 En la figura 4.13 se muestra la gráfica de la función $f(x) = (100/x^2)$ sen $(10/x)$ para x en $[1, 3]$. Al utilizar el algoritmo de la cuadratura adaptativa 4.3 con la tolerancia 10^{-4} para aproximar $\int_1^3 f(x)dx$ obtenemos -1.426014, resultado que tiene una precisión de 1.1×10^{-5}. La aproximación requería aplicar la regla de Simpson con $n = 4$ en los 23 subintervalos cuyos extremos aparecen en el eje horizontal de la figura 4.13. El número total de evaluaciones funcionales que se necesitan en esta aproximación es 93.

Figura 4.13



El valor más grande de h con el que la regla compuesta normal de Simpson da una exactitud de 10^{-4} es $h = \frac{1}{88}$. Esta aplicación requiere 177 evaluaciones de funciones, casi el doble del método adaptativo. ■

CONJUNTO DE EJERCICIOS 4.6

1. Calcule las aproximaciones de la regla de Simpson $S(a, b)$, $S(a, (a + b)/2)$ y $S((a + b)/2, b)$ para las siguientes integrales, y verifique la estimación dada en la fórmula de aproximación.

   a. $\int_1^{1.5} x^2 \ln x \, dx$ b. $\int_0^1 x^2 e^{-x} \, dx$

   c. $\int_0^{0.35} \frac{2}{x^2 - 4} \, dx$ d. $\int_0^{\pi/4} x^2 \operatorname{sen} x \, dx$

   e. $\int_0^{\pi/4} e^{3x} \operatorname{sen} 2x \, dx$ f. $\int_1^{1.6} \frac{2x}{x^2 - 4} \, dx$

   g. $\int_3^{3.5} \frac{x}{\sqrt{x^2 - 4}} \, dx$ h. $\int_0^{\pi/4} (\cos x)^2 \, dx$

2. Use la cuadratura adaptativa para obtener las aproximaciones de las integrales del ejercicio 1 con una exactitud de 10^{-3}. No use software alguno para generar esos resultados.

3. Use la cuadratura adaptativa para aproximar las siguientes integrales con una exactitud de 10^{-5}.

   a. $\int_1^3 e^{2x} \operatorname{sen} 3x \, dx$ b. $\int_1^3 e^{3x} \operatorname{sen} 2x \, dx$

   c. $\int_0^5 [2x \cos(2x) - (x - 2)^2] \, dx$ d. $\int_0^5 [4x \cos(2x) - (x - 2)^2] \, dx$

4. Use la regla compuesta de Simpson con $n = 4, 6, 8, \ldots,$ hasta que las aproximaciones sucesivas de las siguientes integrales concuerden con una exactitud de 10^{-6}. Determine la cantidad de nodos que se requieren. Mediante el algoritmo de la cuadratura adaptativa aproxime la integral con una exactitud de 10^{-6} y cuente el número de nodos. ¿Produjo alguna mejora la cuadratura adaptativa?

   a. $\int_0^\pi x \cos x^2 \, dx$ b. $\int_0^\pi x \operatorname{sen} x^2 \, dx$

   c. $\int_0^\pi x^2 \cos x \, dx$ d. $\int_0^\pi x^2 \operatorname{sen} x \, dx$

5. Dibuje las gráficas de sen $(1/x)$ y de $\cos(1/x)$ en $[0.1, 2]$. Por medio de la cuadratura adaptativa aproxime las integrales

$$\int_{0.1}^2 \operatorname{sen} \frac{1}{x} \, dx \quad \text{y} \quad \int_{0.1}^2 \cos \frac{1}{x} \, dx$$

   con una exactitud de 10^{-3}.

6. Sea $T(a, b)$ y sea $T(a, \frac{a+b}{2}) + T(\frac{a+b}{2}, b)$ las aplicaciones individual y doble de la regla del trapecio a $\int_a^b f(x) \, dx$. Obtenga la relación que existe entre

$$\left| T(a, b) - T\left(a, \frac{a+b}{2}\right) - T\left(\frac{a+b}{2}, b\right) \right|$$

y
$$\left|\int_a^b f(x)\,dx - T\left(a, \frac{a+b}{2}\right) - T\left(\frac{a+b}{2}, b\right)\right|.$$

7. La ecuación diferencial
$$mu''(t) + ku(t) = F_0 \cos \omega t$$
describe un sistema de masa-resorte con una masa m, una constante de resorte k, y sin amortiguamiento. El término $F_0 \cos \omega t$ describe una fuerza externa periódica que se aplica al sistema. La solución de la ecuación cuando el sistema se encuentra inicialmente en reposo ($u'(0) = u(0) = 0$) es
$$u(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)}[\cos \omega t - \cos \omega_0 t], \quad \text{donde} \quad \omega_0 = \sqrt{\frac{k}{m}} \neq \omega.$$
Dibuja la gráfica de u cuando $m = 1$, $k = 9$, $F_0 = 1$, $\omega = 2$ y cuando $t \in [0, 2\pi]$. Aproxime $\int_0^{2\pi} u(t)\,dt$ con una exactitud de 10^{-4}.

8. Si agregamos el término $cu'(t)$ al extremo izquierdo de la ecuación de movimiento del ejercicio 7, la ecuación diferencial resultante describe un sistema de masa-resorte que está amortiguado, con una constante de amortiguamento $c \neq 0$. La solución de esta ecuación cuando el sistema se encuentra inicialmente en reposo es
$$u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + \frac{F_0}{c^2\omega^2 + m^2(\omega_0^2 - \omega^2)^2}[c\omega \operatorname{sen} \omega t + m(\omega_0^2 - \omega^2)\cos \omega t],$$
donde
$$r_1 = \frac{-c + \sqrt{c^2 - 4\omega_0^2 m^2}}{2m} \quad \text{y} \quad r_2 = \frac{-c - \sqrt{c^2 - 4\omega_0^2 m^2}}{2m}.$$

a. Sean $m = 1$, $k = 9$, $F_0 = 1$, $c = 10$, y $\omega = 2$. Determinemos los valores de c_1 y c_2 de modo que $u(0) = u(1) = 0$.

b. Bosqueje la gráfica de $u(t)$ para $t \in [0.2\,\pi]$ y aproxime $\int_0^{2\pi} u(t)\,dt$ con una exactitud de 10^{-4}.

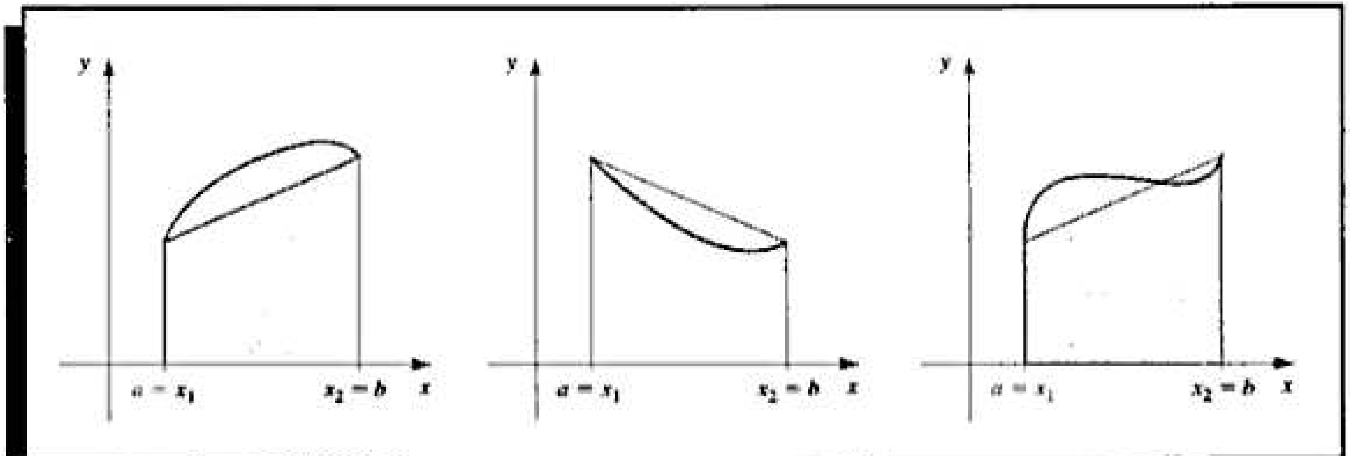
9. El estudio de la difracción de la luz en una apertura rectangular implica el uso de las integrales de Fresnel
$$c(t) = \int_0^t \cos \frac{\pi}{2}w^2\,dw \quad \text{y} \quad s(t) = \int_0^t \operatorname{sen} \frac{\pi}{2}w^2\,dw.$$
Construya una tabla de valores para $c(t)$ y $s(t)$ que tenga una exactitud de 10^{-4} para los valores de $t = 0.1, 0.2, \ldots, 1.0$.

4.7 Cuadratura gaussiana

Las fórmulas de Newton-Cotes de la sección 4.3 se dedujeron integrando los polinomios interpolantes. Puesto que el término de error en el polinomio interpolante de grado n contiene la $(n + 1)$-ésima derivada de la función a aproximar, una fórmula de este tipo será exacta cuando aproxime cualquier polinomio de un grado menor o igual que n.

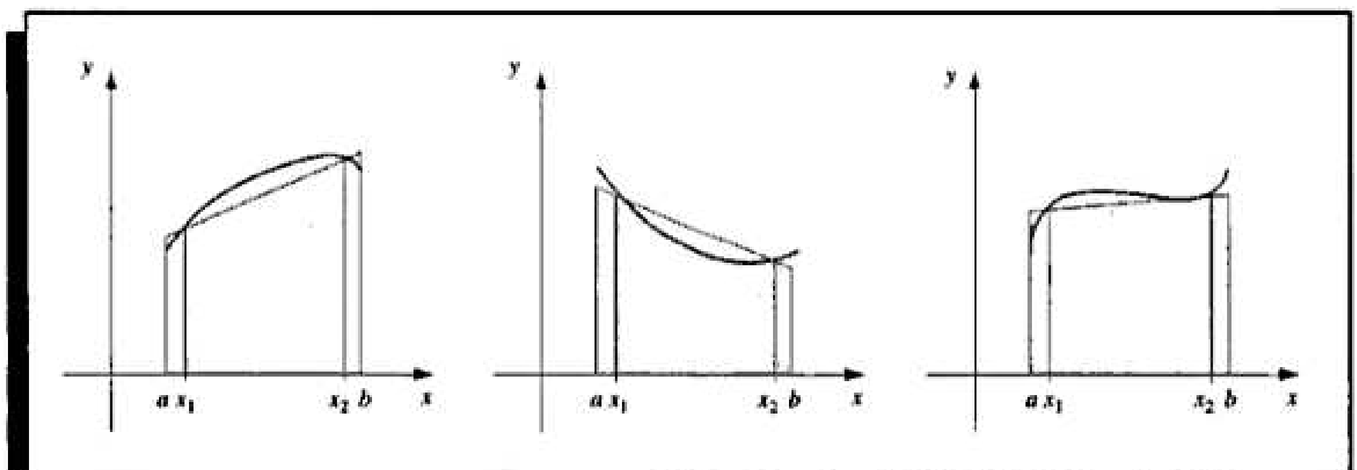
En todas las fórmulas de Newton-Cotes se emplean valores de la función en puntos equidistantes. Esta práctica es adecuada cuando las fórmulas se combinan para formar las reglas compuestas que ya explicamos en la sección 4.4; pero esta restricción puede afectar considerablemente la exactitud de la aproximación. Por ejemplo, tomemos el caso de la regla del trapecio con que se determinan las integrales de las funciones de la figura 4.14.

Figura 4.14



La regla del trapecio aproxima la integral de la función al integrar la función lineal que une los extremos de la gráfica de la función. Pero sin duda ésta no es la mejor línea para aproximar la integral. Las líneas como las que se muestran en la figura 4.15 seguramente producirán, en la generalidad de los casos, mucho mejores aproximaciones.

Figura 4.15



La cuadratura gaussiana selecciona los puntos de la evaluación de manera óptima y no en una forma igualmente espaciada. Se escogen los nodos $x_1, x_2, \ldots, x_n$ en el intervalo

$[a, b]$ y los coeficientes $c_1, c_2, \ldots, c_n$, para reducir en lo posible el error esperado que se obtiene al efectuar la aproximación

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} c_i f(x_i).$$

Si queremos medir esta exactitud, supondremos que la selección óptima de estos valores es la que dé el resultado exacto de la clase más numerosa de polinomios, es decir, la selección que ofrezca el máximo grado de precisión.

En la fórmula de aproximación los coeficientes $c_1, c_2, \ldots, c_n$ son arbitrarios, y los nodos $x_1, x_2, \ldots, x_n$ están restringidos sólo por la especificación de que se encuentren en $[a, b]$, el intervalo de la integración. Esto nos da $2n$ parámetros de donde elegir. Si los coeficientes de un polinomio se consideran parámetros, la clase de polinomios de grado máximo $2n - 1$ también contiene $2n$ parámetros. Así pues, éste es el tipo de polinomios más amplio en que es posible esperar que la fórmula sea exacta. Se puede lograr la exactitud cuando los valores y constantes se seleccionan bien.

Para dar un ejemplo del procedimiento con que se escogen los parámetros apropiados, mostraremos cómo seleccionar los coeficientes y los nodos cuando $n = 2$ y cuando el intervalo de integración es $[-1, 1]$. Después explicaremos el caso más general de una elección arbitraria de los nodos y coeficientes, indicando cómo modificar el método cuando se integra en un intervalo arbitrario.

Supóngase que queremos determinar c_1, c_2, x_1 y x_2 de modo que la fórmula de integración

$$\int_{-1}^{1} f(x)\,dx \approx c_1 f(x_1) + c_2 f(x_2)$$

dé el resultado exacto siempre que $f(x)$ sea un polinomio de grado $2(2) - 1 = 3$ o menor, es decir, cuando

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

para algún conjunto de constantes a_0, a_1, a_2 y a_3. Dado que

$$\int (a_0 + a_1 x + a_2 x^2 + a_3 x^3)\,dx = a_0 \int 1\,dx + a_1 \int x\,dx + a_2 \int x^2\,dx + a_3 \int x^3\,dx,$$

esto equivale a demostrar que la fórmula produce resultados exactos cuando $f(x)$ es 1, x, x^2 y x^3. Por tanto, necesitamos c_1, c_2, x_1 y x_2, de modo que

$$c_1 \cdot 1 + c_2 \cdot 1 = \int_{-1}^{1} 1\,dx = 2, \qquad c_1 \cdot x_1 + c_2 \cdot x_2 = \int_{-1}^{1} x\,dx = 0,$$

$$c_1 \cdot x_1^2 + c_2 \cdot x_2^2 = \int_{-1}^{1} x^2\,dx = \frac{2}{3}, \qquad \text{y} \qquad c_1 \cdot x_1^3 + c_2 \cdot x_2^3 = \int_{-1}^{1} x^3\,dx = 0.$$

Con unas cuantas operaciones algebraicas demostramos que este sistema de ecuaciones tiene solución única

$$c_1 = 1, \qquad c_2 = 1, \qquad x_1 = -\frac{\sqrt{3}}{3} \qquad \text{y} \qquad x_2 = \frac{\sqrt{3}}{3},$$

con que se obtiene la fórmula de aproximación

$$\int_{-1}^{1} f(x)\,dx \approx f\left(\frac{-\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right). \tag{4.41}$$

Esta fórmula tiene un grado de precisión tres, esto es, produce el resultado exacto con cada polinomio de grado tres o menor.

Con esta técnica podríamos determinar los nodos y coeficientes de las fórmulas que proporcionan resultados exactos con los polinomios de grado superior, pero también podemos aplicar un método alterno para obtenerlos más fácilmente. En las secciones 8.2 y 8.3 estudiaremos varios grupos de polinomios ortogonales, que son funciones que tienen la propiedad de que una integral definida del producto de dos de ellos cualesquiera es cero. El conjunto relacionado con nuestro problema es el de los polinomios de Legendre, un conjunto $\{P_0(x), P_1(x), \ldots, P_n(x), \ldots\}$ con las siguientes propiedades:

1. Para cada n, $P_n(x)$ es un polinomio de grado n.
2. $\int_{-1}^{1} P(x)\, P_n(x)\, dx = 0$ siempre que $P(x)$ sea un polinomio de un grado menor que n.

Los primeros polinomios de Legendre son

$$P_0(x) = 1, \qquad P_1(x) = x, \qquad P_2(x) = x^2 - \frac{1}{3},$$

$$P_3(x) = x^3 - \frac{3}{5}x \qquad \text{y} \qquad P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}.$$

Las raíces de estos polinomios son diferentes, se encuentran en el intervalo $(-1, 1)$ tienen simetría con respecto del origen y, lo más importante de todo, es la opción correcta para determinar los parámetros que resuelven nuestro problema.

Los nodos $x_1, x_2, \ldots, x_n$ necesarios para producir una fórmula de la aproximación a la integral, que proporcione resultados exactos para cualquier polinomio de un grado menor que $2n$ son las raíces del polinomio de Legendre de grado n. Esto se establece por medio del siguiente resultado.

Teorema 4.7 Supongamos que $x_1, x_2, \ldots, x_n$ son las raíces del polinomio de Legendre $P_n(x)$ de n-ésimo grado y que para cada $i = 1, 2, \ldots, n$, los números c_i están definidos por

$$c_i = \int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\, dx.$$

Si $P(x)$ es un polinomio cualquiera de un grado menor que $2n$, entonces

$$\int_{-1}^{1} P(x)\, dx = \sum_{i=1}^{n} c_i P(x_i).$$ ■

Demostración Tomemos primero el caso de un polinomio $P(x)$ de un grado menor que n. Reescribimos $P(x)$ como un polinomio de Lagrange de $(n-1)$-ésimo grado, con nodos en las raíces del polinomio de Legendre $P_n(x)$. Esta representación de $P(x)$ es exacta, ya que el término de error contiene la n-ésima derivada de P y esa derivada es cero. Por tanto,

$$\int_{-1}^{1} P(x)\, dx = \int_{-1}^{1} \left[\sum_{i=1}^{n} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} P(x_i) \right] dx$$

$$= \sum_{i=1}^{n} \left[\int_{-1}^{1} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \, dx \right] P(x_i) = \sum_{i=1}^{n} c_i \, P(x_i),$$

con esto verificamos el resultado de los polinomios de un grado menor que n.

Si el polinomio $P(x)$ de un grado menor que $2n$ se divide entre el polinomio de Legendre de n-ésimo grado $P_n(x)$, entonces dos polinomios $Q(x)$ y $R(x)$ de un grado menor que n se producen por medio de

$$P(x) = Q(x) \, P_n(x) + R(x).$$

Ahora recurrimos a la potencia única de los polinomios de Legendre. Primero, el grado del polinomio $Q(x)$ es menor que n; por tanto (de acuerdo con la propiedad 2),

$$\int_{-1}^{1} Q(x) \, P_n(x) \, dx = 0.$$

Después, como x_i es una raíz de $P_n(x)$ para cada $i = 1, 2, \ldots, n$, tenemos

$$P(x_i) = Q(x_i) \, P_n(x_i) + R(x_i) = R(x_i).$$

Finalmente, como $R(x)$ es un polinomio de grado menor que n, el argumento inicial implica que

$$\int_{-1}^{1} R(x) \, dx = \sum_{i=1}^{n} c_i R(x_i).$$

Al combinar estos hechos, verificamos que la fórmula es exacta para el polinomio $P(x)$:

$$\int_{-1}^{1} P(x) \, dx = \int_{-1}^{1} [Q(x) \, P_n(x) + R(x)] \, dx = \int_{-1}^{1} R(x) \, dx = \sum_{i=1}^{n} c_i \, R(x_i) = \sum_{i=1}^{n} c_i \, P(x_i)$$

■ ■ ■

Las constantes c_i necesarias para que la cuadratura funcione, puede generarse a partir de la ecuación del teorema 4.7, pero ambas constantes y las raíces de los polinomios de Legendre se tabulan ampliamente. La tabla 4.11 contiene estos valores para $n = 2, 3, 4$ y 5. Podemos encontrar otras tablas en [StS].

Una integral $\int_a^b f(x)dx$ en un intervalo arbitrario $[a, b]$ se puede transformar, en otra en $[-1, 1]$ usando el cambio de variables (véase Fig. 4.16):

$$t = \frac{2x - a - b}{b - a} \Leftrightarrow x = \frac{1}{2}[(b - a)t + a + b].$$

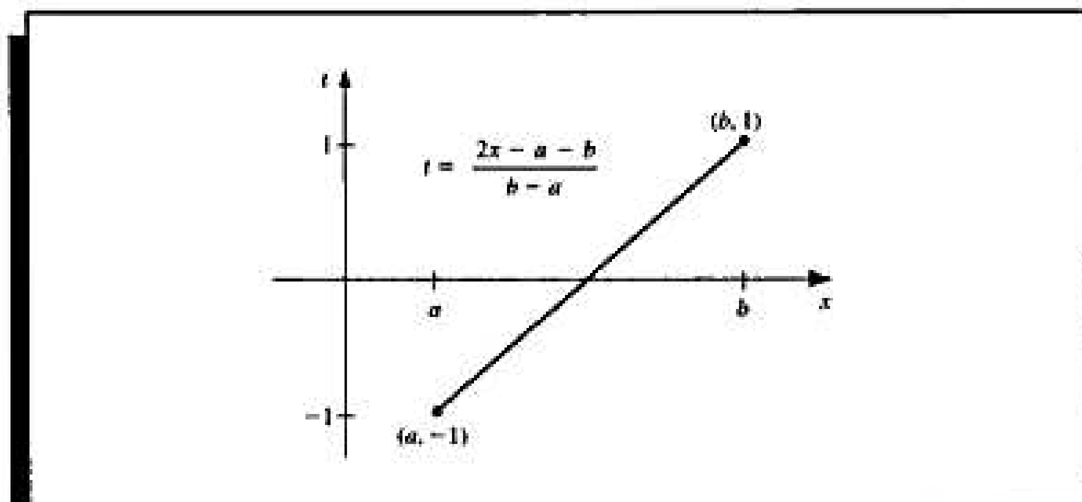
Esto nos permite aplicar la cuadratura gaussiana a cualquier intervalo $[a + b]$, ya que

$$\int_a^b f(x) \, dx = \int_{-1}^{1} f\left(\frac{(b - a)t + (b + a)}{2} \right) \frac{(b - a)}{2} \, dt. \tag{4.42}$$

Tabla 4.11

n	Raíces $r_{n,i}$	Coeficientes $c_{n,i}$
2	0.5773502692	1.0000000000
	−0.5773502692	1.0000000000
3	0.7745966692	0.5555555556
	0.0000000000	0.8888888889
	−0.7745966692	0.5555555556
4	0.8611363116	0.3478548451
	0.3399810436	0.6521451549
	−0.3399810436	0.6521451549
	−0.8611363116	0.3478548451
5	0.9061798459	0.2369268850
	0.5384693101	0.4786286705
	0.0000000000	0.5688888889
	−0.5384693101	0.4786286705
	−0.9061798459	0.2369268850

Figura 4.16



EJEMPLO 1 Consideremos el problema de obtener aproximaciones a $\int_1^{1.5} e^{-x^2}\,dx$. La tabla 4.12 contiene los valores de las fórmulas de Newton-Cotes que vienen en la sección 4.3. El valor exacto de la integral con siete decimales es 0.1093643.

Tabla 4.12

n	0	1	2	3	4
Fórmulas cerradas		0.1183197	0.1093104	0.1093404	0.1093643
Fórmulas abiertas	0.1048057	0.1063473	0.1094116	0.1093971	

El procedimiento de la cuadrat[illegible]ıssiana aplicado a este problema requiere transformar primero la integral en un prob[illegible] cuyo intervalo de integración sea $[-1, 1]$. Al usar la ecuación (4.42) tenemos

$$\int_1^{1.5} e^{-x^2}\,dx = \frac{1}{4}\int_{-1}^{1} e^{(-(t+5)^2/16)}\,dt.$$

Al utilizar los valores de la tabla 4.11, obtenemos mejores aproximaciones de la cuadratura gaussiana en este problema

$n = 2$:

$$\int_1^{1.5} e^{-x^2}\,dx \approx \frac{1}{4}\left[e^{-(5+0.5773502692)^2/16} + e^{-(5-0.5773502692)^2/16}\right] = 0.1094003;$$

$n = 3$:

$$\int_1^{1.5} e^{-x^2}\,dx \approx \frac{1}{4}\left[(0.5555555556)e^{-(5+0.7745966692)^2/16} + (0.8888888889)e^{-(5)^2/16}\right.$$
$$\left. + (0.5555555556)e^{-(5-0.7745966692)^2/16}\right]$$
$$= 0.1093642.$$

Con el fin de facilitar la comparación, en la tabla 4.13 se incluyen los valores obtenidos al aplicar el procedimiento de Rombertg con $n = 4$. ■

Tabla 4.13

0.1183197			
0.1115627	0.1093104		
0.1099114	0.1093610	0.1093643	
0.1095009	0.1093641	0.1093643	0.1093643

CONJUNTO DE EJERCICIOS 4.7

1. Aproxime las siguientes integrales usando la cuadratura gaussiana con $n = 2$ y compare sus resultados con los valores exactos de las integrales.

   a. $\int_1^{1.5} x^2 \ln x\,dx$

   b. $\int_0^1 x^2 e^{-x}\,dx$

   c. $\int_0^{0.35} \frac{2}{x^2-4}\,dx$

   d. $\int_0^{\pi/4} x^2 \operatorname{sen} x\,dx$

   e. $\int_0^{\pi/4} e^{3x} \operatorname{sen} 2x\,dx$

   f. $\int_1^{1.6} \frac{2x}{x^2-4}\,dx$

   g. $\int_3^{3.5} \frac{x}{\sqrt{x^2-4}}\,dx$

   h. $\int_0^{\pi/4} (\cos x)^2\,dx$

2. Repita el ejercicio 1 con $n = 3$.
3. Repita el ejercicio 1 con $n = 4$.
4. Repita el ejercicio 1 con $n = 5$.
5. Determine las constantes a, b, c y d que producirán una fórmula de cuadratura

$$\int_{-1}^{1} f(x)\,dx = af(-1) + bf(1) + cf'(-1) + df'(1)$$

cuyo grado de precisión es 3.

Determine las constantes a, b, c y d que producirán una fórmula de cuadratura

$$\int_{-1}^{1} f(x)\,dx = a f(-1) + b f(0) + c f(1) + d f'(-1) + e f'(1)$$

cuyo grado de precisión es 4.

6. Verifique las entradas o datos para los valores de $n = 2$ y 3 en la tabla 4.11, obteniendo las raíces de los polinomios de Legendre respectivos y usando las ecuaciones anteriores a la tabla para calcular los coeficientes asociados a los valores.
7. Demuestre que la fórmula $Q(P) = \sum_{i=1}^{n} c_i P(x_i)$ no puede generar un grado de precisión mayor que $2n - 1$, sin importar la selección de $c_1, \ldots, c_n$ y $x_1, \ldots, x_n$. [*Sugerencia*: construya un polinomio que tenga una raíz doble en cada una de las x_i.]

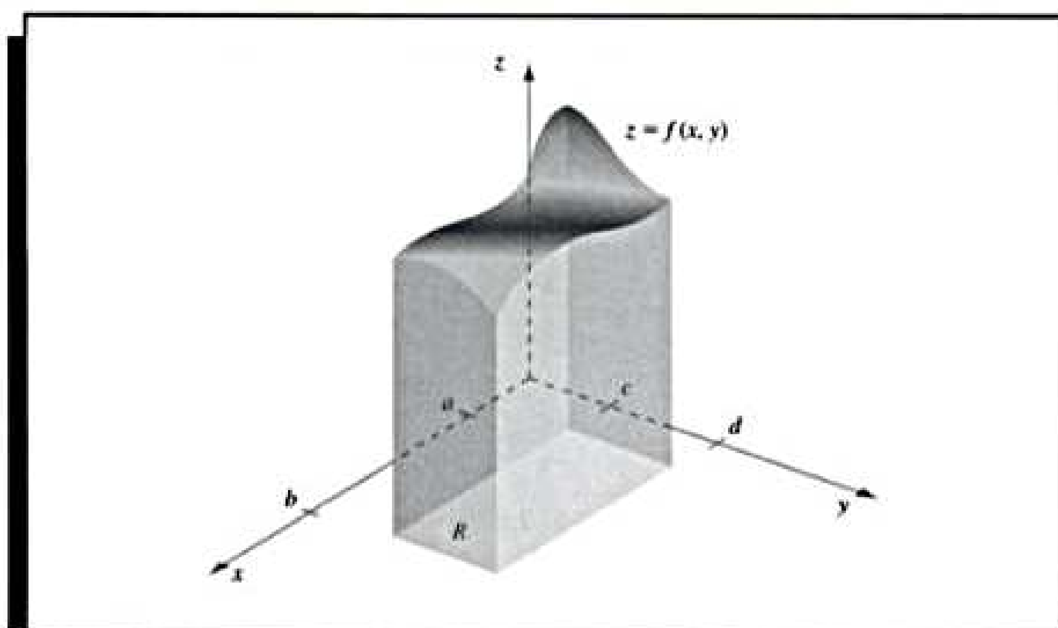
4.8 Integrales múltiples

Podemos modificar abiertamente los métodos explicados en las secciones anteriores y utilizarlos para aproximar integrales múltiples. Consideremos la integral doble

$$\iint_R f(x, y)\,dA,$$

donde R es una región rectangular en el plano: $R = \{(x, y) \mid a \leq x \leq b,\ c \leq y \leq d\}$, para algunas constantes, a, b, c y d (véase Fig. 4.17). Para dar un ejemplo del método de apro-

Figura 4.17



ximación, emplearemos la regla compuesta de Simpson, aunque también podríamos utilizar cualquier otra regla compuesta.

Para aplicar la regla compuesta de Simpson, dividimos la región R fraccionando $[a, b]$ y $[c, d]$ en un número par de intervalos. Para simplificar la notación escogemos los enteros n y m y las particiones $[a, b]$ y $[c, d]$ con los puntos de la red uniformemente espaciados $x_0, x_1, \ldots, x_n$ y $y_0, y_1, \ldots, y_m$, respectivamente. Estas subdivisiones determinan a los tamaños del paso $h = (b - a)/n$ y $k = (d - c)/m$. Al escribir la integral doble como intregal iterada

$$\iint_R f(x, y)\, dA = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx,$$

primero aplicamos la regla compuesta de Simpson para evaluar

$$\int_c^d f(x, y)\, dy,$$

tratando a x como una constante. Sea $y_j = c + jk$ para cada $j = 0, 1, \ldots, m$. Entonces

$$\int_c^d f(x, y)\, dy = \frac{k}{3}\left[f(x, y_0) + 2\sum_{j=1}^{(m/2)-1} f(x, y_{2j}) + 4\sum_{j=1}^{m/2} f(x, y_{2j-1}) + f(x, y_m) \right] - \frac{(d-c)k^4}{180}\frac{\partial^4 f(x, \mu)}{\partial y^4}$$

para alguna μ en (c, d). Por tanto

$$\int_a^b \int_c^d f(x, y)\, dy\, dx = \frac{k}{3}\left[\int_a^b f(x, y_0)\, dx + 2\sum_{j=1}^{(m/2)-1}\int_a^b f(x, y_{2j})\, dx + 4\sum_{j=1}^{m/2}\int_a^b f(x, y_{2j-1})\, dx + \int_a^b f(x, y_m)\, dx \right] - \frac{(d-c)k^4}{180}\int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\, dx.$$

Ahora se aplica en la integrales de esta ecuación la regla compuesta de Simpson. Sea $x_i = a + ih$ para cada $i = 0, 1, \ldots, n$. Entonces, para cada $j = 0, 1, \ldots, m$, tenemos

$$\int_a^b f(x, y_j)\, dx = \frac{h}{3}\left[f(x_0, y_j) + 2\sum_{i=1}^{(n/2)-1} f(x_{2i}, y_j) + 4\sum_{i=1}^{n/2} f(x_{2i-1}, y_j) + f(x_n, y_j) \right] - \frac{(b-a)h^4}{180}\frac{\partial^4 f}{\partial x^4}(\xi_j, y_j),$$

para alguna ξ_j en (a, b). La aproximación resultante tiene la forma:

$$\int_a^b \int_c^d f(x, y)\, dy\, dx \approx \frac{hk}{9} \Bigg\{ \Bigg[f(x_0, y_0) + 2 \sum_{i=1}^{(n/2)-1} f(x_{2i}, y_0)$$

$$+ 4 \sum_{i=1}^{n/2} f(x_{2i-1}, y_0) + f(x_n, y_0) \Bigg]$$

$$+ 2 \Bigg[\sum_{j=1}^{(m/2)-1} f(x_0, y_{2j}) + 2 \sum_{j=1}^{(m/2)-1} \sum_{i=1}^{(n/2)-1} f(x_{2i}, y_{2j})$$

$$+ 4 \sum_{j=1}^{(m/2)-1} \sum_{i=1}^{n/2} f(x_{2i-1}, y_{2j}) + \sum_{j=1}^{(m/2)-1} f(x_n, y_{2j}) \Bigg]$$

$$+ 4 \Bigg[\sum_{j=1}^{(m/2)} f(x_0, y_{2j-1}) + 2 \sum_{j=1}^{m/2} \sum_{i=1}^{(n/2)-1} f(x_{2i}, y_{2j-1})$$

$$+ 4 \sum_{j=1}^{m/2} \sum_{i=1}^{n/2} f(x_{2i-1}, y_{2j-1}) + \sum_{j=1}^{m/2} f(x_n, y_{2j-1}) \Bigg]$$

$$+ \Bigg[f(x_0, y_m) + 2 \sum_{i=1}^{(n/2)-1} f(x_{2i}, y_m) + 4 \sum_{i=1}^{n/2} f(x_{2i-1}, y_m)$$

$$+ f(x_n, y_m) \Bigg] \Bigg\}.$$

El término de error E está dado por

$$E = \frac{-k(b-a)h^4}{540} \left[\frac{\partial^4 f(\xi_0, y_0)}{\partial x^4} + 2 \sum_{j=1}^{(m/2)-1} \frac{\partial^4 f(\xi_{2j}, y_{2j})}{\partial x^4} + 4 \sum_{j=1}^{m/2} \partial^4 \frac{f(\xi_{2j-1}, y_{2j-1})}{\partial x^4} \right.$$

$$\left. + \frac{\partial^4 f(\xi_m, y_m)}{\partial x^4} \right] - \frac{(d-c)k^4}{180} \int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\, dx.$$

Si $\partial^4 f / \partial x^4$ es continua, el teorema del valor intermedio puede aplicarse varias veces para demostrar que la evaluación de las derivadas parciales respecto a x puede ser reemplazada por un valor común y que

$$E = \frac{-k(b-a)h^4}{540} \left[3m \frac{\partial^4 f}{\partial x^4} (\overline{\eta}, \overline{\mu}) \right] - \frac{(d-c)k^4}{180} \int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\, dx,$$

para algunas $(\overline{\eta}, \overline{\mu})$ en R. Si $\partial^4 f / \partial y^4$ también es continua, el teorema del valor medio ponderado de las integrales implica que

$$\int_a^b \frac{\partial^4 f(x, \mu)}{\partial y^4}\, dx = (b-a) \frac{\partial^4 f}{\partial y^4} (\hat{\eta}, \hat{\mu}),$$

para algunas $(\hat{\eta}, \hat{\mu})$ en R. Puesto que $m = (d - c)/k$, el término de error tiene la forma

$$E = \frac{-k(b-a)h^4}{540}\left[3m\,\frac{\partial^4 f}{\partial x^4}(\overline{\eta}, \overline{\mu})\right] - \frac{(d-c)(b-a)}{180}k^4\,\frac{\partial^4 f}{\partial y^4}(\hat{\eta}, \hat{\mu})$$

o

$$E = -\frac{(d-c)(b-a)}{180}\left[h^4\,\frac{\partial^4 f}{\partial x^4}(\overline{\eta}, \overline{\mu}) + k^4\,\frac{\partial^4 f}{\partial y^4}(\hat{\eta}, \hat{\mu})\right],$$

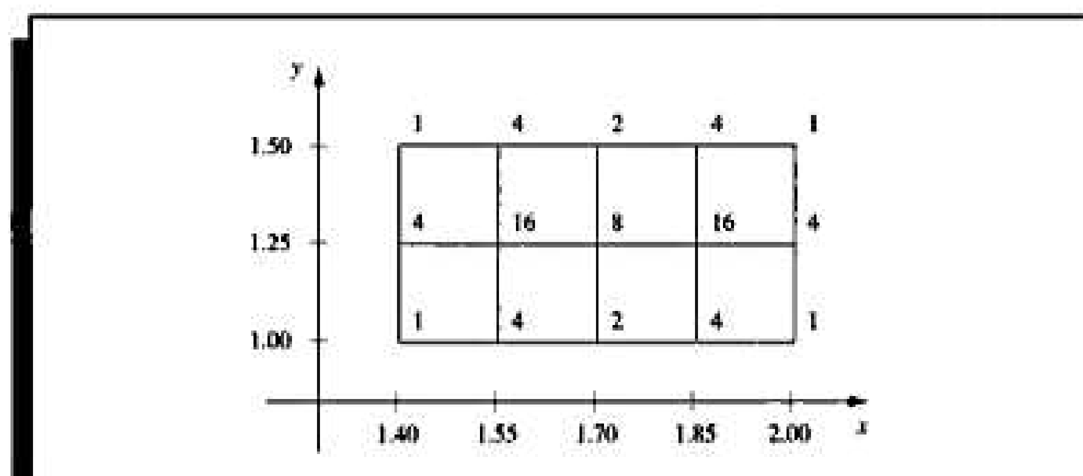
para algunas $(\overline{\eta}, \overline{\mu})$ y $(\hat{\eta}, \hat{\mu})$ en R.

EJEMPLO 1 La regla compuesta de Simpson aplicada para aproximar

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x + 2y)\,dy\,dx,$$

con $n = 4$ y $m = 2$ utiliza los tamaños de paso $h = 0.15$ y $k = 0.25$. En la figura 4.18, se muestra la región de integración R junto con los nodos (x_i, y_j), donde $i = 0, 1, 2, 3, 4$ y $j = 0, 1, 2$ y $w_{i,j}$ que son los coeficientes de $f(x_i, y_j) = \ln(x_i + 2y_j)$ en la suma.

Figura 4.18



La aproximación es

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x + 2y)\,dy\,dx \approx \frac{(0.15)(0.25)}{9}\sum_{i=0}^{4}\sum_{j=0}^{2} w_{i,j}\ln(x_i + 2y_j)$$

$$= 0.4295524387.$$

Puesto que

$$\frac{\partial^4 f}{\partial x^4}(x, y) = \frac{-6}{(x + 2y)^4} \qquad \text{y} \qquad \frac{\partial^4 f}{\partial y^4}(x, y) = \frac{-96}{(x + 2y)^4},$$

y el valor máximo de $\dfrac{1}{(x+2y)^4}$ en R ocurre en (1.4, 1.0), el error estará acotado por

$$
\begin{aligned}
|E| &\leq \frac{(0.5)(0.6)}{180}\left[(0.15)^4 \max_{(x,y)\text{ en } R} \frac{6}{(x+2y)^4} + (0.25)^4 \max_{(x,y)\text{ en } R} \frac{96}{(x+2y)^4}\right] \\
&\leq 4.72 \times 10^{-6}.
\end{aligned}
$$

El valor real de la integral a diez cifras decimales es

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx = 0.4295545265,$$

y, por tanto, la aproximación tiene una exactitud de 2.1×10^{-6}. ■

Podemos aplicar la misma técnica para aproximar las integrales triples y también las integrales superiores de funciones con más de tres variables. La cantidad de evaluaciones funcionales necesarias para la aproximación, es producto del número de las que se requieren cuando aplicamos el método a cada variable.

Si queremos reducir la cantidad de evaluaciones funcionales, en vez de las fórmulas de Newton-Cotes podemos incorporar métodos más eficientes como la cuadratura gaussiana, la integración de Romberg o la cuadratura adaptativa. En el ejemplo siguiente se explica la aplicación de la cuadratura gaussiana a la integral incluida en el ejemplo 1.

EJEMPLO 2 Considere la integral doble del ejemplo 1. Antes de usar la cuadratura gaussiana para aproximarla, transformamos la región de integración

$$R = \{(x, y) \mid 1.4 \leq x \leq 2.0, 1.0 \leq y \leq 1.5\}$$

en

$$\hat{R} = \{(u, v) \mid -1 \leq u \leq 1, -1 \leq v \leq 1\}.$$

Las transformaciones lineales con las que lo logramos son

$$u = \frac{1}{2.0 - 1.4}(2x - 1.4 - 2.0) \quad \text{y} \quad v = \frac{1}{1.5 - 1.0}(2y - 1.0 - 1.5),$$

o, en forma equivalente, $x = 0.3u + 1.7$ y $y = 0.25v + 1.25$. La utilización de este cambio de variables nos da una integral a la cual podemos aplicar la cuadratura gaussiana:

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x+2y)\,dy\,dx = 0.075\int_{-1}^{1}\int_{-1}^{1} \ln(0.3u + 0.5v + 4.2)\,dv\,du.$$

La fórmula de la cuadratura gaussiana para $n = 3$ tanto en u como en v requiere que usemos los nodos

$$u_1 = v_1 = r_{3,2} = 0,\ u_0 = v_0 = r_{3,1} = -0.7745966692,$$

y

$$u_2 = v_2 = r_{3,3} = 0.7745966692.$$

Los pesos asociados son $c_{3,2} = 0.\overline{8}$ y $c_{3,1} = c_{3,3} = 0.\overline{5}$. (Véase la tabla 4.11.) Por tanto,

$$\int_{1.4}^{2.0}\int_{1.0}^{1.5} \ln(x + 2y)\, dy\, dx \approx 0.075 \sum_{i=1}^{3}\sum_{j=1}^{3} c_{3,i}c_{3,j} \ln(0.3r_{3,i} + 0.5r_{3,j} + 4.2)$$
$$= 0.4295545313.$$

Aunque este resultado requiere apenas 9 evaluaciones funcionales en comparación con las 15 que requiere la regla compuesta de Simpson considerada en el ejemplo 1, este resultado tiene una exactitud de 4.8×10^{-9}, en comparación con la exactitud 2.1×10^{-6} del ejemplo 1. ■

El uso de los métodos con que se aproximan las integrales dobles no se limita a las que tienen regiones rectangulares de integración. Los métodos que explicamos anteriormente pueden modificarse para aproximar las integrales dobles de la forma

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx \tag{4.43}$$

o

$$\int_c^d \int_{a(y)}^{b(y)} f(x, y)\, dx\, dy. \tag{4.44}$$

De hecho, también podemos aproximar las integrales en las regiones que no son de este tipo efectuando las particiones de la región adecuadas. (Véase el ejercicio 10.)

Para describir el método que se utiliza al aproximar una integral en la forma

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx,$$

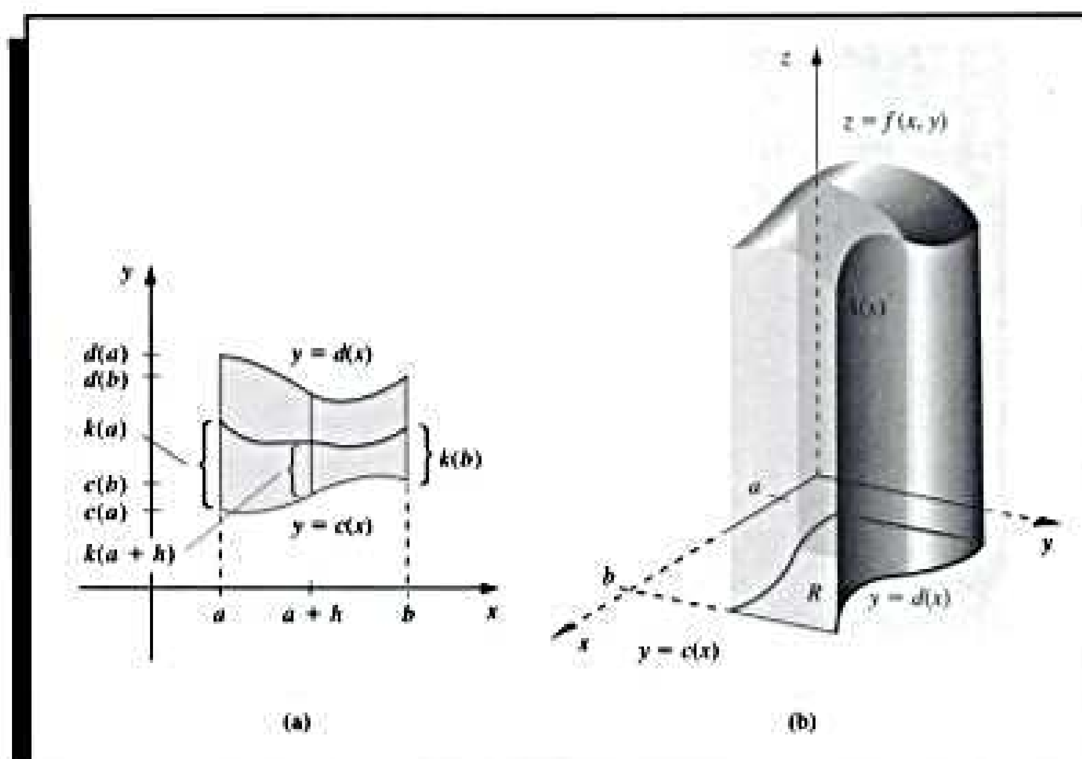
aplicamos la regla de Simpson para integrar respecto a ambas variables. El tamaño del paso de la variable x es $h = (b - a)/2$, pero el tamaño del paso de y varía con x (Véase Fig. 4.19) y se escribe

$$k(x) = \frac{d(x) - c(x)}{2}.$$

En consecuencia,

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx \approx \int_a^b \frac{k(x)}{3}\left[f(x, c(x)) + 4f(x, c(x) + k(x)) + f(x, d(x))\right] dx$$
$$\approx \frac{h}{3}\left\{ \frac{k(a)}{3}\left[f(a, c(a)) + 4f(a, c(a) + k(a)) + f(a, d(a))\right]\right.$$
$$+ \frac{4k(a + h)}{3}\left[f(a + h, c(a + h)) + 4f(a + h, c(a + h)\right.$$
$$\left. + k(a + h)) + f(a + h, d(a + h))\right]$$
$$\left. + \frac{k(b)}{3}\left[f(b, c(b)) + 4f(b, c(b) + k(b)) + f(b, d(b))\right]\right\}.$$

Figura 4.19



El algoritmo 4.4 aplica la regla compuesta de Simpson a una integral de la forma (4.43). Las integrales de la forma (4.44) pueden manejarse de manera semejante.

ALGORITMO 4.4

Integral doble de Simpson

Para aproximar la integral $I = \int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx$:

ENTRADA extremos a, b: enteros positivos pares m, n.

SALIDA aproximación J a I.

Paso 1 Tome $h = (b - a)/n$;
$J_1 = 0$; (*Términos extremos.*)
$J_2 = 0$; (*Términos pares.*)
$J_3 = 0$; (*Términos impares.*)

Paso 2 Para $i = 0, 1, \ldots, n$ haga los pasos 3-8.

Paso 3 Tome $x = a + ih$; (*Método compuesto de Simpson para x.*)
$HX = (d(x) - c(x))/m$;
$K_1 = f(x, c(x)) + f(x, d(x))$; (*Términos extremos.*)
$K_2 = 0$; (*Términos pares.*)
$K_3 = 0$. (*Términos impares.*)

Paso 4 Para $j = 1, 2, \ldots, m - 1$ haga los pasos 5 y 6.

Paso 5 Tome $y = c(x) + jHX$;
$Q = f(x, y)$.

Paso 6 Si j es par, entonces tome $K_2 = K_2 + Q$
si no, tome $K_3 = K_3 + Q$.

Paso 7 Tome $L = (K_1 + 2K_2 + 4K_3)HX/3$.

$$\left(L = \int_{c(x_i)}^{d(x_i)} f(x_i, y)dy \quad \textit{por el método compuesto de Simpson.} \right)$$

Paso 8 Si $i = 0$ o $i = n$ entonces tome $J_1 = J_1 + L$
si no, si i es par entonces tome $J_2 = J_2 + L$
si no, tome $J_3 = J_3 + L$.

Paso 9 Tome $J = h(J_1 + 2J_2 + 4J_3)/3$.

Paso 10 SALIDA (J);
PARAR. ■

Si queremos aplicar la cuadratura gaussiana a

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx,$$

primero debemos transformar, el intervalo $[c(x), d(x)]$ a $[-1, 1]$ para cada x en $[a, b]$ y luego aplicar la cuadratura gaussiana. Esto nos da la fórmula

$$\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx$$
$$\approx \int_a^b \frac{d(x) - c(x)}{2} \sum_{j=1}^{n} c_{n,j} f\left(x, \frac{(d(x) - c(x))r_{n,j} + d(x) + c(x)}{2}\right) dx,$$

donde, como antes, las raíces $r_{n,j}$ y los coeficientes $c_{n,j}$ provienen de la tabla 4.11. Ahora transformamos el intervalo $[a, b]$ en $[-1, 1]$ y usamos la cuadratura gaussiana para aproximar la integral del lado derecho de esta ecuación. Los detalles se incluyen en el algoritmo 4.5.

ALGORITMO 4.5

Integral doble gaussiana

Para aproximar la integral $\int_a^b \int_{c(x)}^{d(x)} f(x, y)\, dy\, dx$:

ENTRADA extremos a, b; enteros positivos m, n.
(Las raíces $r_{i,j}$ y los coeficientes $c_{i,j}$ deben estar disponibles para $\{m, n\}$ y para $1 \le j \le i$.)

SALIDA aproximación J a I.

Paso 1 Tome $h_1 = (b - a)/2$;
$h_2 = (b + a)/2$;
$J = 0$.

Paso 2 Para $i = 1, 2, \ldots, m$ haga los pasos 3-5.

Paso 3 Tome $JX = 0$;
$x = h_1 r_{m,i} + h_2$;
$d_1 = d(x)$;
$c_1 = c(x)$;
$k_1 = (d_1 - c_1)/2$
$k_2 = (d_1 + c_1)/2$.

Paso 4 Para $j = 1, 2, \ldots, n$ haga
tome $y = k_1 r_{n,j} + k_2$;
$Q = f(x, y)$;
$JX = JX + c_{n,j}\, Q$.

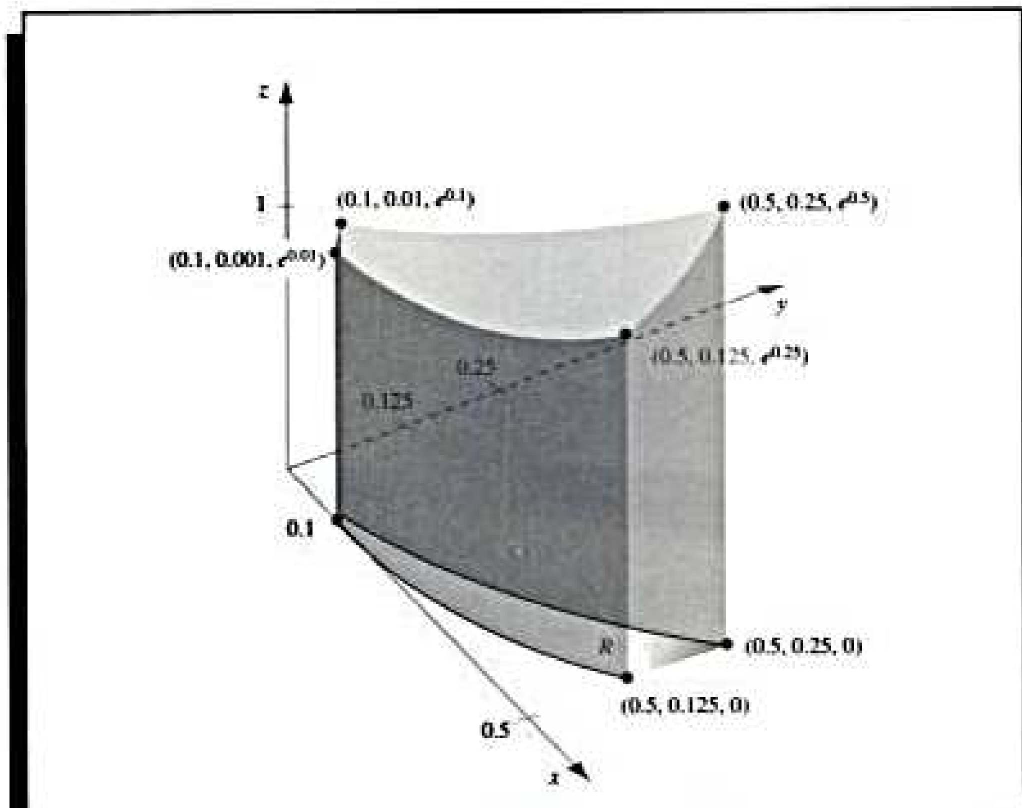
Paso 5 Tome $J = J + c_{m,i} k_1 JX$.

Paso 6 Tome $J = h_1 J$.

Paso 7 SALIDA (J);
PARAR. ■

EJEMPLO 3 El volumen del sólido de la figura 4.20 se aproxima aplicando el algoritmo de la integral doble de Simpson, con $n = m = 10$ a

Figura 4.20



$$\int_{0.1}^{0.5}\int_{x^3}^{x^2} e^{y/x}\, dy\, dx.$$

Esto requiere realizar 121 evaluaciones de la función $f(x, y) = e^{y/x}$ y produce el resultado 0.0333054, el cual se aproxima al volumen del sólido que se muestra en la figura 4.20, cuya exactitud es de casi siete cifras decimales. Si queremos aplicar el algoritmo de la cuadratura gaussiana con $n = m = 5$, necesitamos sólo 25 evaluaciones de la función, obteniéndose así la aproximación 0.03330556611, que tiene una exactitud de 11 cifras decimales. ■

Las integrales triples de la forma

$$\int_a^b \int_{c(x)}^{d(x)} \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz\, dy\, dx$$

(véase Fig. 421) se aproximan de manera similar. Debido a la cantidad de cálculos que se requieren, la cuadratura gaussiana es el método indicado. En el algoritmo 4.6 se aplica este procedimiento.

ALGORITMO 4.6

Integral triple gaussiana

Para aproximar la integral $\int_a^b \int_{c(x)}^{d(x)} \int_{\alpha(x,y)}^{\beta(x,y)} f(x, y, z)\, dz\, dy\, dx$:

ENTRADA extremos a, b; enteros positivos m, n, p.
(Las raíces $r_{i,j}$ y los coeficientes $c_{i,j}$ deben estar disponibles para i = máx $\{n, m, p\}$ y para $1 \le j \le i$.)

SALIDA aproximación J a I.

Paso 1 Tome $h_1 = (b - a)/2$;
$h_2 = (b + a)/2$;
$J = 0$.

Paso 2 Para $i = 1, 2, \ldots, m$ haga los pasos 3-8.

Paso 3 Tome $JX = 0$;
$x = h_1 r_{m,i} + h_2$;
$d_1 = d(x)$;
$c_1 = c(x)$;
$k_1 = (d_1 - c_1)/2$;
$k_2 = (d_1 + c_1)/2$.

Paso 4 Para $j = 1, 2, \ldots, n$ haga los pasos 5-7.

Paso 5 Tome $JY = 0$;
$y = k_1 r_{n,j} + k_2$;
$\beta_1 = \beta(x, y)$;
$\alpha_1 = \alpha(x, y)$;
$l_1 = (\beta_1 - \alpha_1)/2$;
$l_2 = (\beta_1 + \alpha_1)/2$.

Paso 6 Para $k = 1, 2, \ldots, p$ haga
tome $z = l_1 r_{p,k} + l_2$;
$Q = f(x, y, z)$;
$JY = JY + c_{p,k} Q$.

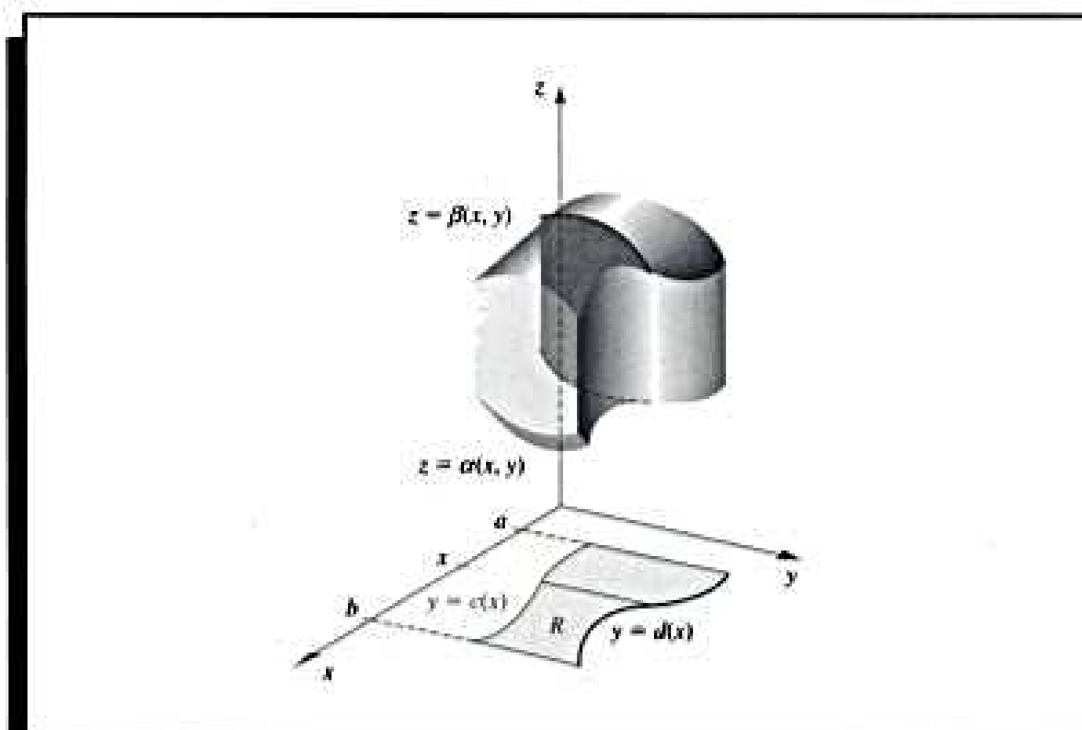
Paso 7 Tome $JX = JX + c_{n,j} l_1 JY$.

Paso 8 Tome $J = J + c_{m,i} k_1 JX$.

Paso 9 Tome $J = h_1 J$.

Paso 10 SALIDA (J);
PARAR. ■

Figura 4.21



El ejemplo siguiente requiere la evaluación de cuatro integrales triples.

EJEMPLO 4 El centro de una masa de una región sólida D con la función de densidad σ se halla en

$$(\bar{x}, \bar{y}, \bar{z}) = \left(\frac{M_{yz}}{M}, \frac{M_{xz}}{M}, \frac{M_{xy}}{M}\right),$$

donde

$$M_{yz} = \iiint_D x\sigma(x, y, z)\, dV, \quad M_{xz} = \iiint_D y\sigma(x, y, z)\, dV$$

y

$$M_{xy} = \iiint_D z\sigma(x, y, z)\, dV$$

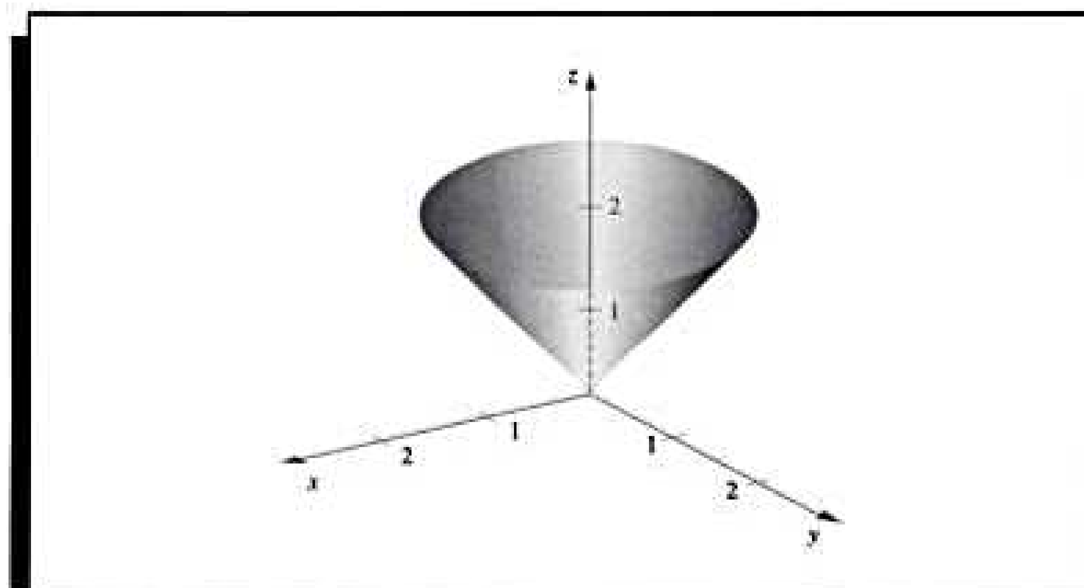
son los momentos alrededor de los planos coordenados y donde

$$M = \iiint_D \sigma(x, y, z)\, dV$$

es la masa. El sólido de la figura 4.22 está acotado por la parte superior del cono que divide el vértice $z^2 = x^2 + y^2$ y el plano $z = 2$ y tiene la función de densidad dada por

$$\sigma(x, y, z) = \sqrt{x^2 + y^2}.$$

Figura 4.22



Al aplicar el algoritmo de la integral triple gaussiana 4.6 con $n = m = p = 5$ se requiere realizar 125 evaluaciones de función por integral, y se obtienen las siguientes aproximaciones:

$$M = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} \sqrt{x^2 + y^2}\, dz\, dy\, dx$$

$$= 4 \int_{0}^{2} \int_{0}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} \sqrt{x^2 + y^2}\, dz\, dy\, dx \approx 8.37504476,$$

$$M_{yz} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} x\sqrt{x^2 + y^2}\, dz\, dy\, dx \approx -5.55111512 \times 10^{-17},$$

$$M_{xz} = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} \int_{\sqrt{x^2+y^2}}^{2} y\sqrt{x^2 + y^2}\, dz\, dy\, dx \approx -8.01513675 \times 10^{-17},$$

$$M_{xy} = \int_{-2}^{2}\int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}}\int_{\sqrt{x^2+y^2}}^{2} z\sqrt{x^2+y^2}\,dz\,dy\,dx \approx 13.40038156.$$

Esto significa que la ubicación aproximada del centro de masa es

$$(\bar{x}, \bar{y}, \bar{z}) = (0, 0, 1.60003701).$$

Por medio de la evaluación directa de las integrales, podemos demostrar que el centro de masa se encuentra en (0, 0, 1.6). ■

CONJUNTO DE EJERCICIOS 4.8

1. Use el algoritmo 4.4 con $n = m = 4$ para aproximar las siguientes integrales dobles y después compare los resultados con los valores exactos.

a. $\int_{2.1}^{2.5}\int_{1.2}^{1.4} xy^2\,dy\,dx$ **b.** $\int_{0}^{0.5}\int_{0}^{0.5} e^{y-x}\,dy\,dx$

c. $\int_{2}^{2.2}\int_{x}^{2x} (x^2+y^3)\,dy\,dx$ **d.** $\int_{1}^{1.5}\int_{0}^{x} (x^2+\sqrt{y})\,dy\,dx$

2. Calcule los valores más pequeños cuando $n = m$, de modo que pueda emplear el algoritmo 4.4 para aproximar las integrales del ejercicio 1 con una exactitud de 10^{-6} del valor real.

3. Use el algoritmo 4.4 con (i) $n = 4$, $m = 8$, (ii) $n = 8$, $m = 4$ y (iii) $n = m = 6$ para aproximar las siguientes integrales dobles y luego compare los resultados con las respuestas exactas.

a. $\int_{0}^{\pi/4}\int_{\operatorname{sen} x}^{\cos x} (2y \operatorname{sen} x + \cos^2 x)\,dy\,dx$ **b.** $\int_{1}^{e}\int_{1}^{x} \ln xy\,dy\,dx$

c. $\int_{0}^{1}\int_{x}^{2x} (x^2+y^3)\,dy\,dx$ **d.** $\int_{0}^{1}\int_{x}^{2x} (y^2+x^3)\,dy\,dx$

e. $\int_{0}^{\pi}\int_{0}^{x} \cos x\,dy\,dx$ **f.** $\int_{0}^{\pi}\int_{0}^{x} \cos y\,dy\,dx$

g. $\int_{0}^{\pi/4}\int_{0}^{\operatorname{sen} x} \frac{1}{\sqrt{1-y^2}}\,dy\,dx$ **h.** $\int_{-\pi}^{3\pi/2}\int_{0}^{2\pi} (y \operatorname{sen} x + x \cos y)\,dy\,dx$

4. Obtenga los valores más pequeños cuando $n = m$, de manera que pueda emplear el algoritmo 4.4 para aproximar las integrales del ejercicio 3 con una exactitud de 10^{-6} del valor real.

5. Use el algoritmo 4.5 cuando $n = m = 2$ para aproximar las integrales del ejercicio 1, y después compare los resultados con los que obtuvo en el ejercicio 1.

6. Calcule los valores más pequeños de $n = m$, de modo que pueda usar el algoritmo 4.5 para aproximar las integrales del ejercicio 1 con una exactitud de 10^{-6}. No vaya más allá de $n = m = 5$. Compare la cantidad de evaluaciones funcionales requeridas con la cantidad requerida en el ejercicio 2.

7. Use el algoritmo 4.5 con (i) $n = m = 3$, (ii) $n = 3$, $m = 4$, (iii) $n = 4$, $m = 3$ y (iv) $n = m = 4$ para aproximar las integrales del ejercicio 3.

8. Use el algoritmo 4.5 con $n = m = 5$ para aproximar las integrales del ejercicio 3. Compare la cantidad de evaluaciones de funciones requeridas con la cantidad que se necesita en el ejercicio 4.

9. Use el algoritmo 4.4 con $n = m = 14$ y el algoritmo 4.5 con $n = m = 4$ para aproximar

$$\iint_R e^{-(x+y)}\, dA,$$

para la región R en el plano acotado por las curvas $y = x^2$ y $y = \sqrt{x}$.

10. Use el algoritmo 4.4 para aproximar

$$\iint_R \sqrt{xy + y^2}\, dA,$$

donde R es la región del plano acotada por las líneas $x + y = 6$, $3y - x = 2$ y $3x - y = 2$. Primero, divida R en las regiones R_1 y R_2 en las que el algoritmo 4.4 pueda aplicarse. Utilice $n = m = 6$ tanto en R_1 como R_2.

11. Una lámina plana es una hoja delgada de masa uniformemente distribuida. Si σ es una función que describe la densidad de una lámina que tiene la forma de una región R en el plano xy, entonces el centro de masa de la lámina $(\bar{x}, \bar{y})$ está definido por

$$x = \frac{\iint_R x\sigma(x, y)\, dA}{\iint_R \sigma(x, y)\, dA} \qquad \bar{y} = \frac{\iint_R y\sigma(x, y)\, dA}{\iint_R \sigma(x, y)\, dA}.$$

Use el algoritmo 4.4 con $n = m = 14$ para encontrar el centro de masa de la lámina descrita por $R = \{(x, y) \mid 0 \le x \le 1, 0 \le y \le \sqrt{1 - x^2}\}$ con la función de densidad $\sigma(x, y)e^{-(x^2+y^2)}$. Compare la aproximación con el resultado exacto.

12. Repita el ejercicio 11 empleando el algoritmo 4.5 con $n = m = 5$.

13. El área de la superficie descrita por $z = f(x, y)$ para (x, y) en R está dada por

$$\iint_R \sqrt{[f_x(x, y)]^2 + [f_y(x, y)]^2 + 1}\, dA.$$

Use el algoritmo 4.4 con $n = m = 8$ para obtener una aproximación al área de la superficie en el hemisferio $x^2 + y^2 + z^2 = 9$, $z \ge 0$ que se encuentra arriba de la región en el plano descrito por $R = \{(x, y) \mid 0 \le x \le 1, 0 \le y \le 1\}$.

14. Repita el ejercicio 13 aplicando ahora el algoritmo 4.5 con $n = m = 4$.

15. Use el algoritmo 4.6 con $n = m = 2$ para aproximar las siguientes integrales triples, y después compare los resultados con los valores exactos.

a. $\int_0^1 \int_1^2 \int_0^{0.5} e^{x+y+z}\, dz\, dy\, dx$

b. $\int_0^1 \int_x^1 \int_0^y y^2 z\, dz\, dy\, dx$

c. $\int_0^1 \int_{x^2}^x \int_{x-y}^{x+y} y\, dz\, dy\, dx$

d. $\int_0^1 \int_{x^2}^x \int_{x-y}^{x+y} z\, dz\, dy\, dx$

e. $\int_0^\pi \int_0^x \int_0^{xy} \frac{1}{y} \operatorname{sen} \frac{z}{y}\, dz\, dy\, dx$

f. $\int_0^1 \int_0^1 \int_{-xy}^{xy} e^{x^2+y^2}\, dz\, dy\, dx$

16. Repita el ejercicio 15 usando $n = m = p = 3$.

17. Repita el ejercicio 15 usando $n = m = p = 4$ y $n = m = p = 5$.

18. Use el algoritmo 4.6 con $n = m = p = 4$ para aproximar

$$\iiint_S xy \operatorname{sen}(yz)\, dV,$$

donde S es el sólido acotado por los planos coordenados y por los planos $x = \pi$, $y = \pi/2$, $z = \pi/3$. Compare esta aproximación con el resultado exacto.

19. Use el algoritmo 4.6 con $n = m = p = 5$ para aproximar

$$\iiint_S \sqrt{xyz}\, dV,$$

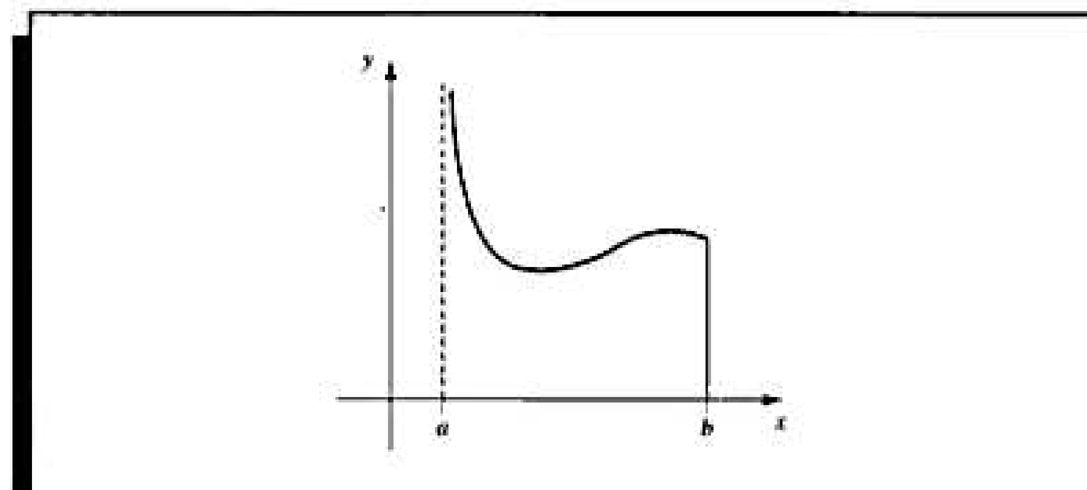
donde S es la región del primer octante acotada por el cilindro $x^2 + y^2 = 4$, la esfera $x^2 + y^2 + z^2 = 4$, y el plano $x + y + z = 8$. ¿Cuántas evaluaciones de la función se requieren en la aproximación?

4.9 Integrales impropias

Las integrales impropias se producen cuando el concepto de integración se extiende a un intervalo de integración donde la función no está acotada, o a un intervalo con uno o más extremos infinitos. En ambos casos, es preciso modificar las reglas normales de la aproximación de la integral.

Primero consideraremos la situación en que el integrando no está acotado en el extremo izquierdo del intervalo de la integración, como se observa en la figura 4.23. Después mostraremos que, con un manejo adecuado, podemos reducir las otras integrales apropiadas a problemas de este tipo.

Figura 4.23



En cálculo se demuestra que la integral impropia con una singularidad en el extremo izquierdo,

$$\int_a^b \frac{dx}{(x-a)^p},$$

converge si y sólo si $0 < p < 1$, y en este caso definimos

$$\int_a^b \frac{dx}{(x-a)^p} = \frac{(b-a)^{1-p}}{1-p}.$$

Si f es una función que puede escribirse en la forma

$$f(x) = \frac{g(x)}{(x-a)^p},$$

donde $0 < p < 1$ y g es continua en $[a, b]$, entonces la integral impropia

$$\int_a^b f(x)\, dx$$

también existe. Aproximaremos esta integral por medio de la regla compuesta de Simpson. Si $g \in C^5[a, b]$ podremos construir el cuarto polinomio de Taylor $P_4(x)$, para g alrededor de a,

$$P_4(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \frac{g'''(a)}{3!}(x-a)^3 + \frac{g^{(4)}(a)}{4!}(x-a)^4,$$

y escribir

$$\int_a^b f(x)\, dx = \int_a^b \frac{g(x) - P_4(x)}{(x-a)^p}\, dx + \int_a^b \frac{P_4(x)}{(x-a)^p}\, dx. \tag{4.45}$$

Como $P(x)$ es un polinomio, podemos determinar exactamente el valor de

$$\int_a^b \frac{P_4(x)}{(x-a)^p}\, dx = \sum_{k=0}^{4} \int_a^b \frac{g^k(a)}{k!}(x-a)^{k-p}\, dx = \sum_{k=0}^{4} \frac{g^{(k)}(a)}{k!\,(k+1-p)}(b-a)^{k+1-p}. \tag{4.46}$$

Por lo general esta es la parte dominante de la aproximación, especialmente cuando el polinomio de Taylor $P_4(x)$ concuerda estrechamente con $g(x)$ en todo el intervalo $[a, b]$.

Así pues, para aproximar la integral de f tenemos que agregar este valor a la aproximación de

$$\int_a^b \frac{g(x) - P_4(x)}{(x-a)^p}\, dx.$$

Para determinar esto, primero defina

$$G(x) = \begin{cases} \dfrac{g(x) - P_4(x)}{(x-a)^p}, & \text{si } a < x \le b, \\ 0, & \text{si } x = a. \end{cases}$$

Como $0 < p < 1$ y como $P_4^{(k)}(a)$ concuerdan con $g^{(k)}(a)$ para cada $k = 0, 1, 2, 3, 4$, tenemos $G \in C^4[a, b]$. Ello significa que podemos aplicar la regla compuesta de Simpson para aproximar la integral de G en $[a, b]$. Al agregar esta aproximación al valor de la ecuación (4.46), obtenemos una aproximación a la integral impropia de f en $[a, b]$, con la exactitud de la aproximación de la regla compuesta de Simpson.

EJEMPLO 1 Aplicamos la regla compuesta de Simpson $h = 0.25$ para aproximar los valores de la integral imprópia

$$\int_0^1 \frac{e^x}{\sqrt{x}}\,dx.$$

Dado que el cuarto polinomio de Taylor para e^x alrededor de $x = 0$ es

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24},$$

tenemos

$$\int_0^1 \frac{P_4(x)}{\sqrt{x}}\,dx = \int_0^1 \left(x^{-1/2} + x^{1/2} + \frac{1}{2}x^{3/2} + \frac{1}{6}x^{5/2} + \frac{1}{24}x^{7/2}\right)dx$$

$$= \lim_{M\to 0+}\left[2x^{1/2} + \frac{2}{3}x^{3/2} + \frac{1}{5}x^{5/2} + \frac{1}{21}x^{7/2} + \frac{1}{108}x^{9/2}\right]_M^1$$

$$=2 + \frac{2}{3} + \frac{1}{5} + \frac{1}{21} + \frac{1}{108} \approx 2.9235450.$$

La tabla 4.14 contiene los valores aproximados de

$$G(x) = \begin{cases} \dfrac{1}{\sqrt{x}}(e^x - P_4(x)), & \text{cuando } 0 < x \le 1,\\ 0, & \text{cuando } x = 0. \end{cases}$$

Tabla 4.14

x	$G(x)$
0.00	0
0.25	0.0000170
0.50	0.0004013
0.75	0.0026026
1.00	0.0099485

Al aplicar la regla compuesta de Simpson a G usando estos datos, obtenemos

$$\int_0^1 G(x)\,dx \approx \frac{0.25}{3}[0 + 4(0.0000170) + 2(0.0004013) + 4(0.0026026) + 0.0099485] = 0.0017691.$$

Por tanto,

$$\int_0^1 \frac{e^x}{\sqrt{x}}\,dx \approx 2.9235450 + 0.0017691 = 2.9253141.$$

Este resultado tiene una exactitud que corresponde a la de la aproximación obtenida mediante la regla compuesta de Simpson para la función G. Puesto que $|G^{(4)}(x)| < 1$ en $[0, 1]$, el error está acotado por

$$\frac{1-0}{180}(0.25)^4 = 0.0000217. \qquad \blacksquare$$

Para aproximar la integral impropia con una singularidad en el extremo derecho, simplemente aplicamos el método anterior, pero desarrollando en términos del extremo derecho b en lugar del extremo izquierdo a. También podríamos realizar la sustitución

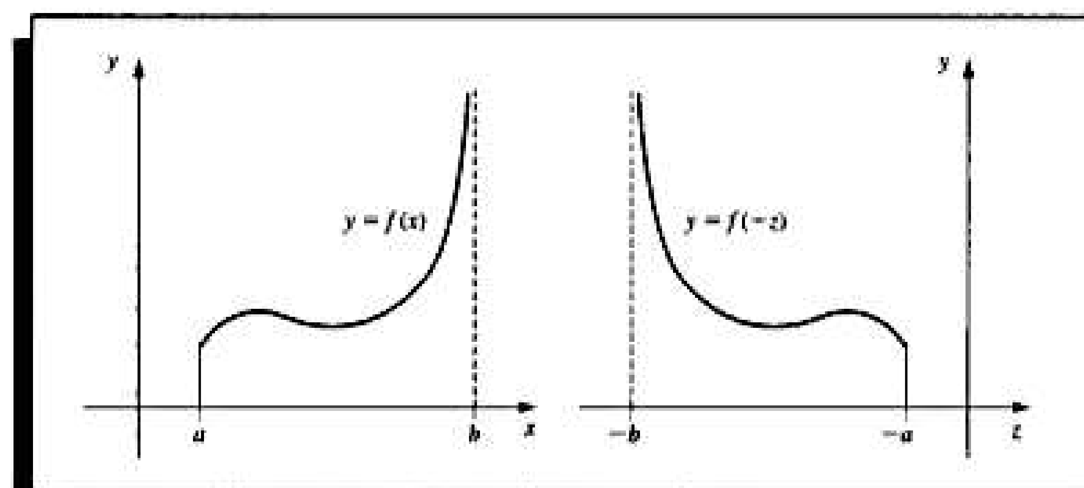
$$z = -x, \qquad dz = -dx$$

para que la integral impropia adquiera la forma

$$\int_a^b f(x)\,dx = \int_{-b}^{-a} f(-z)\,dz, \tag{4.47}$$

que tiene su singularidad en el extremo izquierdo. (Véase la fig. 4.24.)

Figura 4.24



Las integrales impropias con singularidades (por ejemplo, en c), donde $a < c < b$, se tratan como la suma de integrales impropias con singularidades en los extremos, puesto que

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

El otro tipo de integrales impropias contiene límites de integración infinitos. La integral básica de esta clase presenta la forma,

$$\int_a^\infty \frac{1}{x^p}\,dx,$$

para $p > 1$. La convertimos en una integral con singularidad de extremo izquierdo en cero al realizar la sustitución de integración

$$t = x^{-1}, \qquad dt = -x^{-2}\,dx, \qquad \text{entonces} \qquad dx = -x^2\,dt = -t^{-2}\,dt.$$

Por tanto,

$$\int_a^\infty \frac{1}{x^p}\,dx = \int_{1/a}^0 -\frac{t^p}{t^2}\,dt = \int_0^{1/a} \frac{1}{t^{2-p}}\,dt.$$

De modo similar, el cambio de variable $t = x^{-1}$ convierte la integral impropia $\int_a^\infty f(x)\,dx$ en otra que tiene una singularidad de extremo izquierdo en cero:

$$\int_a^\infty f(x)\,dx = \int_0^{1/a} t^{-2} f\left(\frac{1}{t}\right) dt. \tag{4.48}$$

Podemos aproximarla utilizando la fórmula de la cuadratura del tipo que se describió en páginas anteriores.

EJEMPLO 2 Para aproximar el valor de la integral impropia

$$I = \int_1^\infty x^{-3/2} \operatorname{sen} \frac{1}{x}\,dx,$$

hacemos el cambio de variable $t = x^{-1}$ para obtener

$$I = \int_0^1 t^{-1/2} \operatorname{sen} t\,dt.$$

El cuarto polinomio de Taylor, $P_4(t)$, para sen t alrededor de 0 es

$$P_4(t) = t - \frac{1}{6}t^3,$$

así

$$G(t) = \begin{cases} \dfrac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}, & \text{de modo que } 0 < t \le 1 \\ 0, & \text{si } t = 0 \end{cases}$$

está en $C^4[0, 1]$, y tenemos

$$\begin{aligned} I &= \int_0^1 t^{1/2} - \frac{1}{6}t^{5/2}\,dt + \int_0^1 \frac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt \\ &= \left[\frac{2}{3}t^{3/2} - \frac{1}{21}t^{7/2}\right]_0^1 + \int_0^1 \frac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt \\ &= 0.61904761 + \int_0^1 \frac{\operatorname{sen} t - t + \frac{1}{6}t^3}{t^{1/2}}\,dt. \end{aligned}$$

Al aplicar la regla compuesta de Simpson con $n = 16$ a los enteros restantes, obtenemos

$$I = 0.0014890097 + 0.61904761 = 0.62053661,$$

con una exactitud de 4.0×10^{-8}. ■

CONJUNTO DE EJERCICIOS 4.9

1. Aplique la regla compuesta de Simpson y los valores dados de n para aproximar las siguientes integrales impropias:

a. $\int_0^1 x^{-1/4} \operatorname{sen} x\, dx, \quad n = 4$

b. $\int_0^1 \frac{e^{2x}}{\sqrt[5]{x^2}}\, dx, \quad n = 6$

c. $\int_1^2 \frac{\ln x}{(x-1)^{1/5}}\, dx, \quad n = 8$

d. $\int_0^1 \frac{\cos 2x}{x^{1/3}}\, dx, \quad n = 6$

2. Use la regla compuesta de Simpson y los valores dados de n para aproximar las siguientes integrales impropias:

a. $\int_0^1 \frac{e^{-x}}{\sqrt{1-x}}\, dx, \quad n = 6$

b. $\int_0^2 \frac{xe^x}{\sqrt[3]{(x-1)^2}}\, dx, \quad n = 8$

3. Utilice la transformación $t = x^{-1}$, y después la regla compuesta de Simpson y los valores dados de n para aproximar las siguientes integrales impropias:

a. $\int_1^\infty \frac{1}{x^2+9}\, dx, \quad n = 4$

b. $\int_1^\infty \frac{1}{1+x^4}\, dx, \quad n = 4$

c. $\int_1^\infty \frac{\cos x}{x^3}\, dx, \quad n = 6$

d. $\int_1^\infty x^{-4} \operatorname{sen} x\, dx, \quad n = 6$

4. La integral impropia $\int_0^\infty f(x)\, dx$ no puede convertirse en una integral con límites finitos por medio de la sustitución $t = 1/x$ porque el límite en cero se vuelve infinito. El problema se resuelve escribiendo primero $\int_0^\infty f(x)\, dx = \int_0^1 f(x)\, dx + \int_1^\infty f(x)\, dx$. Aplique este método para aproximar las siguientes integrales impropias con una exactitud de 10^{-6}.

a. $\int_0^\infty \frac{1}{1+x^4}\, dx$

b. $\int_0^\infty \frac{1}{(1+x^2)^3}\, dx$

5. Suponga que un cuerpo de masa m se desplaza verticalmente hacia arriba comenzando en la superficie de la Tierra. Si prescindimos de toda la resistencia menos la gravedad, la velocidad de escape v está dada por

$$v^2 = 2gR \int_1^\infty z^{-2}\, dz, \qquad \text{donde } z = \frac{x}{R},$$

$R = 3960$ mi es el radio de la Tierra y $g = 0.00609$ mi/s^2 es la fuerza de gravedad en la superficie de la Tierra. Aproxime la velocidad de escape v.

6. Los polinomios de Laguerre $\{L_0(x), L_1(x), \ldots\}$ forman un conjunto ortogonal en $[0, \infty)$ y satisfacen $\int_0^\infty e^{-x} L_i(x) L_j(x)\, dx = 0$, para $i \neq j$. Véase la sección 8. 2.) El polinomio $L_n(x)$ tiene n ceros distintos $x_1, x_2, \ldots, x_n$ en $[0, \infty)$. Sea

$$c_{n,i} = \int_0^\infty e^{-x} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\, dx.$$

Muestre que la fórmula de cuadratura

$$\int_0^\infty f(x)e^{-x}\, dx = \sum_{i=1}^{n} c_{n,i} f(x_i)$$

tiene grado de precisión $2n - 1$. (*Sugerencia*: Siga los pasos de la demostración del teorema 4.7.)

7. Los polinomios de Laguerre $L_0(x) = 1$, $L_1(x) = 1 - x$, $L_2(x) = x_2 - 4x + 2$ y $L_3(x) = -x^3 + 9x^2 - 18x + 6$ fueron obtenidos en el ejercicio 11 de la sección 8.2. Como mostramos en el ejercicio 6, estos polinomios son útiles para aproximar integrales de la forma

$$\int_0^\infty e^{-x} f(x)\, dx = 0.$$

   a. Deduzca la fórmula de cuadratura usando $n = 2$ y los ceros de $L_2(x)$.

   b. Deduzca la fórmula de cuadratura usando $n = 3$ y los ceros de $L_3(x)$.

8. Use las fórmulas de cuadratura obtenidas en el ejercicio 7 para aproximar la integral

$$\int_0^\infty \sqrt{x}e^{-x}\, dx.$$

9. Use las fórmulas de cuadratura obtenidas en el ejercicio 7 para aproximar la integral

$$\int_{-\infty}^\infty \frac{1}{1 + x^2}\, dx.$$

4.10 Reseña de métodos y software

En este capítulo estudiamos la aproximación de integrales de funciones de una, dos o tres variables y la aproximación de las derivadas de una función con una sola variable real.

Explicamos las reglas del punto medio, del trapecio y de Simpson para describir los métodos y el análisis del error de los métodos de cuadratura. La regla compuesta de Simpson es fácil de usar y da aproximaciones exactas, a menos que la función oscile en un subintervalo del intervalo o de integración. La cuadratura adaptativa se puede aplicar si se sospecha que la función presenta un comportamiento oscilatorio. Usamos la cuadratura gaussiana para reducir en lo posible el número de nodos y para aumentar el grado de precisión. Estudiamos la integración de Romberg para aprovechar la regla y la extrapolación compuesta del trapecio, de fácil aplicación.

La mayor parte de los programas de computación que sirven para integrar una función de una sola variable real, se basan en el método adaptativo o en fórmulas gaussianas extremadamente exactas. La integración cautelosa de Romberg es un método adaptativo que incluye una verificación para asegurarse de que el integrando tiene un comportamiento uniforme en los subintervalos del intervalo de la integración. Este método se usa con gran éxito en las bibliotecas de programas. Por lo general, las integrales múltiples se aproximan ampliando buenos métodos adaptativos a dimensiones superiores. También recomendamos la cuadratura de tipo gaussiano para disminuir la cantidad de evaluaciones de las funciones.

Las principales rutinas de las bibliotecas IMSL y NAG se basan en *QUADPACK: un paquete para integración automática* de R. Piessens, E. de Doncker-Kapenga, C. W. Uberhuber, y D. K. Kahaner publicada por Springer-Verlag en 1983 [PDUK]. Las rutinas también están disponibles como programas de dominio público en http://www.netlib.org/quadpack.

La biblioteca IMSL contiene la función QDAG, que es un esquema de integración adaptativa que se basa en la regla de 21 puntos de Gaussian-Kronrod y que utiliza una regla gaussiana de 10 puntos en la estimación del error. La regla gaussiana usa los diez puntos $x_1, \ldots, x_{10}$ y los pesos $w_1, \ldots, w_{10}$ para producir la fórmula de cuadratura $\sum_{i=1}^{10} w_i f(x_i)$ con que se aproxima $\int_a^b f(x)\, dx$. Los puntos adicionales $x_{11}, \ldots, x_{21}$, y los pesos nuevos $v_1, \ldots, v_{21}$, se emplean posteriormente en la fórmula de Kronrod $\sum_{i=1}^{21} v_i f(x_i)$. Para eliminar el error, se comparan los resultados de ambas fórmulas. La ventaja de utilizar $x_1, \ldots, x_{10}$ en cada fórmula consiste en que f ha de evaluarse sólo en 21 puntos. Si aplicáramos las re-

glas gaussianas independientes de 10 y 21 puntos, se necesitarían 31 evaluaciones de funciones. Este procedimiento admite singularidades de los extremos en el integrando.

Otras subrutinas de IMSL son QDAGS que permiten singularidades en los extremos; QDAGP que permite singularidades especificadas por el usuario; QDAGI, que permite intervalos infinitos de integración y QDNG, que es un procedimiento no adaptativo para funciones uniformes. La subrutina TWODQ usa las reglas de Gauss-Kronrod para integrar una función de dos variables. También existe una subrutina QAND para usar la cuadratura gaussiana e integrar una función de n variables en n intervalos de la forma $[a_i, b_i]$.

La biblioteca NAG incluye las subrutinas D01AJF con las que se calcula la integral de f en el intervalo $[a, b]$ aplicando un método adaptativo que se basa en la cuadratura gaussiana y que utiliza la regla de 10 puntos de Gauss y la regla de 21 puntos de Kronrod. La subrutina D01AHF sirve para aproximar $\int_a^b f(x)\,dx$ con una familia de fórmulas de tipo gaussiano que se basan en 1, 3, 5, 7, 15, 31, 63, 127 y 255 nodos. Estas reglas interrelacionadas de alta precisión se deben a Patterson [Pat] y se emplean en forma adaptativa. La subrutina D01GBF se utiliza con integrales múltiples y D01GAF aproxima una integral cuando sólo se dan puntos de datos y no la función f. NAG contiene muchas otras subrutinas para aproximar integrales.

La llamada a la función de Maple

```
>int(f,x=a..b);
```

calcula la integral definida $\int_a^b f(x)\,dx$. El método numérico que utiliza Maple usa rutinas que maneja la singularidad y luego la cuadratura de Clenshaw-Curtis, que se describe en [CC]. Si esto falla, debido a la presencia de singularidades dentro o cerca del intervalo, entonces se aplica un método de cuadratura adaptativo con exponencial doble. La fórmula adaptativa de Newton-Cotes se puede aplicar especificando la opción `_NCrule` en la llamada a función de Maple.

```
>int(f,x=a..b,digits,_NCrule);
```

El método trata de alcanzar una tolerancia de error relativa de $0.5 \times 10^{(1-\text{Digits})}$, donde Digits es la variable de Maple que especifica la cantidad de dígitos de redondeo que Maple emplea en el cálculo numérico. El valor estándar de Digits es 10, pero podemos transformarlo en cualquier entero positivo n mediante el comando `Digits:=n;` El comando QUAD de MATLAB aproxima la integral definida $\int_a^b f(x)\,dx$ usando una regla adaptativa de Simpson y QUAD8 aproxima la integral definida usando una regla adaptativa de Newton-Cotes de ocho espacios.

Aunque la diferenciación numérica es inestable, se requieren las fórmulas de aproximación de las derivadas para resolver las ecuaciones diferenciales. La biblioteca NAG contiene la subrutina D04AAF para la diferenciación numérica de una función de una variable real que permite diferenciar hasta la derivada. La función DERIV de IMSL emplea un cambio adaptativo en el tamaño del paso de las diferencias finitas para aproximar una derivada f en x con una tolerancia determinada. IMSL incluye, además, la subrutina QDDER con que se calculan las derivadas de una función definida en un conjunto de puntos usando la interpolación cuadrática. Ambos paquetes permiten diferencias e integran los trazadores cúbicos interpolantes que se construyen con las subrutinas mencionadas en la sección 3.4.

Al lector que desee mayor información sobre la integración numérica, le recomendamos los libros de Engels [E] y de Davis y Rabinowitz [DR]. Y para mayor información sobre la cuadratura gaussiana puede consultar a Stroud y Secrest [StS]. Entre los libros que tratan de las integrales múltiples se encuentran los de Stroud [Stro] y el libro reciente de Sloan y Joe [SJ].

CAPÍTULO

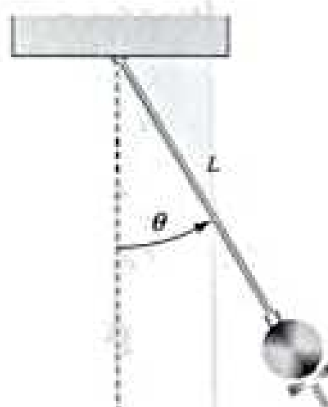
5

Problemas de valor inicial para ecuaciones diferenciales ordinarias

. . .

Con algunas suposiciones simplificadoras, podemos describir el movimiento de un péndulo por medio de la ecuación diferencial de segundo orden

$$\frac{d^2\theta}{dt^2} + \frac{g}{L} \operatorname{sen} \theta = 0,$$



donde L es la longitud del péndulo, $g = 32.17$ pies/s² es la constante gravitacional de la Tierra y θ es el ángulo que forma el péndulo en la posición vertical o de equilibrio. Si además especificamos la posición del péndulo al momento de iniciar el movimiento $\theta(t_0) = \theta'_0$, y su velocidad en ese momento $\theta'(t_0) = \theta'_0$ tendremos lo que se conoce con el nombre de *problema de valor inicial*.

Para simplificar este problema a uno lineal de valor inicial, para valores pequeños de θ, podemos emplear la aproximación $\theta \approx$ sen θ:

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\,\theta = 0, \qquad \theta(t_0) = \theta_0, \quad \theta'(t_0) = \theta'_0.$$

Podemos resolver este problema por medio de un método estándar de ecuaciones diferenciales. Para valores mayores de θ, hay que utilizar métodos de aproximación. En el ejercicio 6 de la sección 5.9, se incluye un problema de este tipo.

En cualquier libro sobre ecuaciones diferenciales encontrará explicaciones amplias acerca de varios métodos para obtener explícitamente soluciones a los problemas de valor inicial de primer orden. Pero, en la práctica, pocos de los problemas que se presentan en el estudio de los fenómenos físicos pueden resolverse con exactitud.

En la primera parte del capítulo estudiaremos cómo aproximar la solución $y(t)$ a un problema de la forma

$$\frac{dy}{dt} = f(t, y), \quad \text{para } a \leq t \leq b,$$

sujeto a la condición inicial

$$y(a) = \alpha.$$

Más adelante en el capítulo, trataremos de la extensión de estos métodos a un sistema de ecuaciones diferenciales de primer orden de la forma

$$\frac{dy_1}{dt} = f_1(t, y_1, y_2, \ldots, y_n),$$

$$\frac{dy_2}{dt} = f_2(t, y_1, y_2, \ldots, y_n),$$

$$\frac{dy_n}{dt} = f_n(t, y_1, y_2, \ldots, y_n),$$

para $a \leq t \leq b$, sujeto a las condiciones iniciales

$$y_1(a) = \alpha_1, \quad y_2(a) = \alpha_2 \quad, \ldots, \quad y_n(a) = \alpha_n,$$

y la relación de un sistema de este tipo con el problema general de valor inicial de n-ésimo orden de la forma

$$y^{(n)} = f(t, y, y', y'', \ldots, y^{(n-1)}),$$

para $a \leq t \leq b$, sujeto a las condiciones iniciales

$$y(a) = \alpha_1, \quad y'(a) = \alpha_2, \quad \ldots, \quad y^{n-1}(a) = \alpha_n.$$

5.1 Teoría elemental de los problemas de valor inicial

Las ecuaciones diferenciales sirven para modelar problemas de ciencias e ingeniería que requieren el cambio de una variable respecto a otra. En la mayor parte de ellos hay que resolver un *problema de valor inicial*, es decir, resolver una ecuación diferencial que satisface una condición inicial dada.

En la generalidad de las situaciones de la vida real, la ecuación diferencial que modela el problema resulta demasiado complicada para resolverla con exactitud, por lo que se recurre a dos procedimientos para aproximar la solución. El primero consiste en simplificar la ecuación diferencial de modo que podamos resolverla exactamente y utilizar después la solución de la ecuación simplificada para aproximar la solución de la ecuación original. El segundo, que examinaremos en este capítulo, se vale de métodos para aproximar la solución del problema original. Este procedimiento es el que se emplea por lo regular, pues los métodos de aproximación dan resultados más exactos y una información realista sobre el error.

Los métodos que veremos en este capítulo no producen una aproximación continua a la solución del problema de valor inicial. Por el contrario, se obtienen las aproximaciones en algunos puntos específicos y, a menudo, igualmente espaciados. Si se requieren valores intermedios, se utiliza un método de interpolación, que generalmente es el de Hermite.

Antes de estudiar los métodos para aproximar los problemas de valor inicial, necesitamos algunas definiciones y resultados de la teoría de las ecuaciones diferenciales ordinarias. Los problemas de valor inicial que planteamos al observar los fenómenos físicos sólo suelen aproximar la situación general, por lo cual necesitamos saber si cambios pequeños en el enunciado del problema introducen cambios igualmente pequeños en la solución. Esto también es importante por la aparición del error de redondeo cuando se recurre a métodos numéricos.

Definición 5.1 Se dice que una función $f(t, y)$ satisface una **condición Lipschitz** en la variable y en un conjunto $D \subset \mathbb{R}^2$ si existe una constante $L > 0$ con la propiedad de que

$$|f(t, y_1) - f(t, y_2)| \leq L\,|y_1 - y_2|,$$

siempre que $(t, y_1), (t, y_2) \in D$. A la constante L se le llama constante de Lipschitz para f. ∎

EJEMPLO 1 Si $D = \{(t, y) \mid 1 \leq t \leq 2, -3 \leq y \leq 4\}$ y $f(t, y) = t|y|$, entonces para cada par de puntos (t, y_1) y (t, y_2) en D tenemos

$$|f(t, y_1) - f(t, y_2)| = |t|y_1| - t|y_2|| = |t|\,||y_1| - |y_2|| \leq 2|y_1 - y_2|.$$

Por tanto, f satisface una condición de Lipschitz en D en la variable y con la constante 2 de Lipschitz. En este problema el valor más pequeño de la constante de Lipschitz que se puede obtener es $L = 2$, así que, por ejemplo,

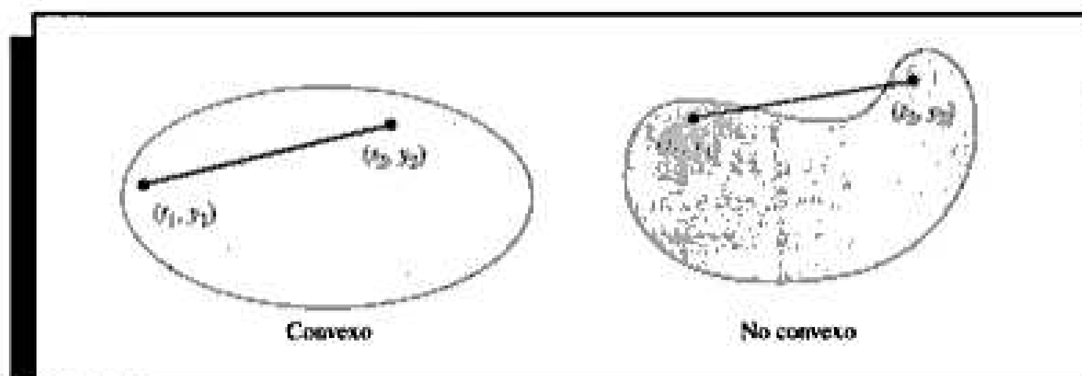
$$|f(2, 1) - f(2, 0)| = |2 - 0| = 2|1 - 0|.$$ ∎

Definición 5.2 Se dice que un conjunto $D \subset \mathbb{R}^2$ es **convexo**, si siempre que (t_1, y_1) y (t_2, y_2) pertenecen a D, el punto $((1 - \lambda)t_1 + \lambda t_2, (1 - \lambda)y_1 + \lambda y_2)$ también pertenece a D para cada λ en $[0, 1]$. ∎

En términos geométricos, la definición 5.2 establece que un conjunto es convexo a condición de que, siempre que dos puntos pertenezcan a él, el segmento de recta entero en-

tre los puntos también pertenezca al conjunto (véase Fig. 5.1). Los conjuntos que consideraremos en este capítulo normalmente son de la forma $D = \{(t, y) \mid a \leq t \leq b, -\infty < y < \infty\}$ para algunas constantes a y b. Es fácil verificar (véase el ejercicio 5) que son convexos.

Figura 5.1



Teorema 5.3 Supongamos que $f(t, y)$ está definida en un conjunto convexo $D \subset \mathbb{R}^2$. Si existe una constante $L > 0$ con

$$\left|\frac{\partial f}{\partial y}(t, y)\right| \leq L, \quad \text{para toda } (t, y) \in D \tag{5.1}$$

entonces f satisface una condición de Lipschitz en D en la variable y con la constante L de Lipschitz. ■

En el ejercicio 4 se da la demostración del teorema 5.3; se parece a la demostración del resultado correspondiente de las funciones de una variable que explicamos en el ejercicio 25 de la sección 1.1.

Como veremos en el siguiente teorema, a menudo es muy importante determinar si la función que interviene en un problema de valor inicial satisface la condición de Lipschitz en su segunda variable, y casi siempre es más fácil aplicar la condición (5.1) que la definición. No obstante, conviene aclarar que el teorema 5.3 sólo da condiciones suficientes para que una condición de Lipschitz sea válida; un reanálisis del ejemplo 1 demostrará que dichas condiciones *no son necesarias* en absoluto.

Así, la función del ejemplo 1 satisface una condición de Lipschitz, pero la derivada respecto de y ni siquiera existe cuando $y = 0$.

El siguiente teorema es una versión del teorema fundamental de existencia y unicidad de las ecuaciones diferenciales ordinarias de primer grado. Aunque podemos demostrarlo reduciendo un poco la hipótesis, esta versión es suficiente para este capítulo. (La demostración del teorema, más o menos en esta forma, se da en [BiR, pp. 142-155].)

Teorema 5.4 Supongamos que $D = \{(t, y) \mid a \leq t \leq b, -\infty < y < \infty\}$ y que $f(t, y)$ es continua en D. Si f satisface una condición de Lipschitz en D en la variable y, entonces el problema de valor inicial

$$y'(t) = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

tiene un solución única $y(t)$ para $a \leq t \leq b$. ■

EJEMPLO 2 Consideremos el problema del valor inicial

$$y' = 1 + t\,\text{sen}(ty), \quad 0 \leq t \leq 2, \quad y(0) = 0.$$

Si mantenemos la constante t y si aplicamos el teorema del valor medio a la función

$$f(t, y) = 1 + t\,\text{sen}(ty),$$

comprobaremos que, siempre que $y_1 < y_2$, existe un número ξ en (y_1, y_2) con

$$\frac{f(t, y_2) - f(t, y_1)}{y_2 - y_1} = \frac{\partial}{\partial y} f(t, \xi) = t^2 \cos(\xi t).$$

Por tanto,

$$|f(t, y_2) - f(t, y_1)| = |y_2 - y_1|\,|t^2 \cos(\xi t)| \leq 4|y_2 - y_1|$$

y f satisface una condición de Lipschitz en la variable y con la constante de Lipschitz $L = 4$. Y como además $f(t, y)$ es continua cuando $0 \leq t \leq 2$ y $-\infty < y < \infty$, el teorema 5.4 implica que este problema de valor inicial tiene una solución única.

Si el lector terminó un curso de ecuaciones diferenciales, posiblemente intentará encontrar la solución exacta del problema. ■

Ahora que, en cierto modo, hemos contestado la pregunta de que cuándo los problemas de valor inicial tienen soluciones únicas, podemos abordar la cuestión planteada anteriormente en la sección:

> ¿Cómo sabemos si un problema tiene la propiedad de que pequeños cambios o perturbaciones del planteamiento del problema ocasionen cambios igualmente pequeños en la solución?

Como de costumbre, primero necesitamos formular una definición operacional que exprese este concepto.

Definición 5.5 Se dice que el problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \qquad a \leq t \leq b, \qquad y(a) = \alpha, \tag{5.2}$$

es un **problema bien planteado** si:

1. El problema tiene una solución única, $y(t)$;
2. Para cualquier $\varepsilon > 0$, existe una constante positiva $k(\varepsilon)$ con la propiedad de que siempre que $|\varepsilon_0| < \varepsilon$ y $\delta(t)$ es continua con $|\delta(t)| < \varepsilon$ en [a, b], existe una solución única, $z(t)$, al problema,

$$\frac{dz}{dt} = f(t, z), + \delta(t), \qquad a \leq t \leq b, \qquad z(a) = \alpha + \varepsilon_0, \tag{5.3}$$

con

$$|z(t) - y(t)| < k(\varepsilon)\varepsilon, \qquad \text{para toda } a \leq t \leq b.$$ ■

Al problema especificado por la ecuación (5.3) se le llama **problema perturbado** asociado al problema original (5.2), y supone la posibilidad de que haya un error $\delta(t)$ en la formulación de la ecuación diferencial y también que la condición inicial contenga un error ε_0.

Los métodos numéricos siempre se ocuparán de resolver un problema perturbado, porque cualquier error de redondeo introducido en la representación altera el problema original. Y si este último no está bien planteado, existen pocas razones para suponer que la solución numérica de un problema de este tipo se aproxima con exactitud a la solución del problema original.

En el teorema siguiente se especifican las condiciones que garantizan el buen planteamiento de un problema de valor inicial. La demostración del teorema se incluye en [BiR, pp. 142-147].

Teorema 5.6 Suponga que $D = \{(t, y) \mid a \leq t \leq b \text{ y } -\infty < y < \infty\}$. Si f es continua y satisface la condición de Lipschitz en la variable y en el conjunto D, entonces el problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha$$

es bien planteado. ■

EJEMPLO 3 Sea $D = \{(t, y) \mid 0 \leq t \leq 1 -\infty < y < \infty\}$, consideremos el problema de valor inicial

$$\frac{dy}{dt} = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5. \tag{5.4}$$

Dado que

$$\left|\frac{\partial(y - t^2 + 1)}{\partial y}\right| = |1| = 1,$$

el teorema 5.3 implica que $f(t, y) = y - t^2 + 1$ satisface una condición de Lipschitz en D con la constante 1 de Lipschitz. Puesto que f es continua en D, el teorema 5.6 implica que el problema es bien planteado.

Para verificar esto directamente, consideremos el problema perturbado

$$\frac{dz}{dt} = z - t^2 + 1 + \delta, \qquad 0 \leq t \leq 2, \qquad z(0) = 0.5 + \varepsilon_0, \tag{5.5}$$

donde δ y ε_0 son constantes. Las soluciones de las ecuaciones (5.4) y (5.5) son

$$y(t) = (t + 1)^2 - 0.5e^t \qquad \text{y} \qquad z(t) = (t + 1)^2 + (\delta + \epsilon_0 - 0.5)e^t - \delta,$$

respectivamente. Es fácil verificar que si, $|\delta| < \varepsilon$ y $|\varepsilon_0| < \varepsilon$, entonces

$$|y(t) - z(t)| = |(\delta + \varepsilon_0)e^t - \delta| \leq |\delta + \varepsilon_0|e^2 + |\delta| \leq (2e^2 + 1)\varepsilon,$$

para toda t. Así, el problema (5.4) está bien planteado, con $k(\varepsilon) = 2e^2 + 1$ para todo $\varepsilon > 0$. ■

Podemos usar Maple para resolver muchos problemas de valor inicial. Consideremos el problema

$$\frac{dy}{dt} = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5.$$

Para definir la ecuación diferencial, introducimos

```
>deq:=D(y)(t)=y(t)-t*t+1;
```

y la condición inicial

```
>init:=y(0)=0.5;
```

Los nombres `deq` e `init` los elige el usuario. El comando para resolver el valor inicial es

```
>deqsol:=dsolve({deq,init},y(t));
```

La respuesta es

$$\text{deqsol:} = y(t) = 1 + t^2 + 2t - \frac{1}{2}e^t$$

Para usar la solución y obtener $y(1.5)$, introducimos

```
>q:=rhs(deqsol);evalf(subs(t=1.5,q));
```

con el resultado 4.009155465.

La función `rhs` sirve para asignar la solución del problema de valor inicial a la función q, que después evaluamos en $t = 1.5$. La función `dsolve` puede fallar, si no se obtiene una solución explícita al problema de valor inicial. Por ejemplo, el comando

```
>deqsol2:=dsolve({D(y)(t)=1+t*sen(t*y(t)),y(0)=0},y(t));
```

no tiene éxito porque no es posible encontrar una solución explícita. En este caso hay que usar un método numérico.

CONJUNTO DE EJERCICIOS 5.1

1. Use el teorema 5.4 para demostrar que los siguientes problemas de valor inicial tienen una solución única, y encuentre dicha solución.
   a. $y' = y\cos t,\quad 0 \le t \le 1,\qquad y(0) = 1.$
   b. $y' = \frac{2}{t}y + t^2e^t,\quad 1 \le t \le 2,\qquad y(1) = 0.$
   c. $y' = -\frac{2}{t}y + t^2e^t,\quad 1 \le t \le 2,\qquad y(1) = \sqrt{2}e.$
   d. $y' = \dfrac{4t^3y}{1+t^4},\qquad 0 \le t \le 1,\qquad y(0) = 1.$
2. Para cada elección de $f(t, y)$ dada en la parte (a)-(d):
   (i) ¿Satisface f una condición de Lipschitz en $D = \{(t, y)\ |\ 0 \le t \le 1, -\infty < y < \infty\}$?
   (ii) ¿Puede usarse el teorema 5.6 para demostrar que el problema de valor inicial

   $$y' = f(t, y),\qquad 0 \le t \le 1,\qquad y(0) = 1,$$

   es bien planteado?
   a. $f(t, y) = t^2y + 1$ b. $f(t, y) = ty$
   c. $f(t, y) = 1 - y$ d. $f(t, y) = -ty + \dfrac{4t}{y}$

3. En los siguientes problemas de valor inicial, demuestre que la ecuación dada define implícitamente una solución. Aproxime $y(2)$ con el método de Newton.

a. $y' = -\dfrac{y^3 + y}{(3y^2 + 1)t}$, $1 \le t \le 2$, $y(1) = 1$; $y^3t + yt = 2$

b. $y' = -\dfrac{y \cos t + 2te^y}{\operatorname{sen} t + t^2e^y + 2}$, $1 \le t \le 2$, $y(1) = 0$; $y \operatorname{sen} t + t^2e^y + 2y = 1$

4. Mediante la aplicación del teorema del valor medio a $f(t, y)$, demuestre el teorema 5.3 conservando fijo a t.

5. Demuestre que, para cualesquiera constantes a y b, el conjunto $D = \{(t, y) \mid a \le t \le b, -\infty < y < \infty\}$ es convexo.

6. Suponga que la perturbación $\delta(t)$ es proporcional a t, es decir, que $\delta(t) = \delta t$ para alguna constante δ. Demuestre directamente que los problemas de valor inicial siguientes están bien planteados.

a. $y' = 1 - y$, $0 \le t \le 2$, $y(0) = 0$

b. $y' = t + y$, $0 \le t \le 2$, $y(0) = -1$

c. $y' = \dfrac{2}{t}y + t^2e^t$, $1 \le t \le 2$, $y(1) = 0$

d. $y' = -\dfrac{2}{t}y + t^2e^t$, $1 \le t \le 2$, $y(1) = \sqrt{2}e$

7. El *método de Picard* para resolver el problema de valor inicial

$$y' = f(t, y), \qquad a \le t \le b, \qquad y(a) = \alpha,$$

se describe así: sea $y_0(t) = \alpha$ para cada t en $[a, b]$. Defina una sucesión $\{y_k(t)\}$ de las funciones por medio de

$$y_k(t) = \alpha + \int_a^t f(\tau, y_{k-1}(\tau))\, d\tau, \qquad k = 1, 2, \ldots$$

a. Integre $y' = f(t, y(t))$ y use la condición inicial para deducir el método de Picard.

b. Genere $y_0(t)$, $y_1(t)$, $y_2(t)$ y $y_3(t)$ para el problema de valor inicial

$$y' = -y + t + 1, \qquad 0 \le t \le 1, \qquad y(0) = 1.$$

c. Compare el resultado de la parte (b) con la serie de Maclaurin de la solución real $y(t) = t + e^{-t}$.

5.2 Método de Euler

En esta sección estudiaremos el método de Euler. Aunque rara vez se emplea en la práctica, la simplicidad de su deducción sirve para ejemplificar las técnicas con que se desarrollan algunos de los métodos más avanzados, sin el álgebra tan engorrosa que acompaña a tales desarrollos.

Este método tiene por objeto obtener una aproximación de un problema bien planteado de valor inicial

$$\frac{dy}{dt} = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha. \tag{5.6}$$

En la práctica, no se obtendrá una aproximación continua a la solución $y(t)$; por el contrario, se generarán aproximaciones a esa solución en varios valores, llamados **puntos de red**, en el intervalo $[a, b]$. Una vez obtenida la aproximación en los puntos, podemos obtener por interpolación la solución aproximada en otros puntos del intervalo.

En primer lugar, estipulamos que los puntos de red tienen una distribución uniforme en todo el intervalo $[a, b]$. Garantizamos esta condición al seleccionar un entero positivo N y los puntos de red

$$t_i = a + ih, \quad \text{para cada } i = 0, 1, 2, \ldots, N.$$

La distancia común entre los puntos $h = (b - a)/N$ recibe el nombre de **tamaño de paso**.

Utilizaremos el teorema de Taylor para derivar el método de Euler. Supongamos que $y(t)$, la solución única de la ecuación (5.6), tiene dos derivadas continuas en $[a, b]$, de modo que para cada $i = 0, 1, 2, \ldots, N - 1$,

$$y(t_{i+1}) = y(t_i) + (t_{i+1} - t_i)y'(t_i) + \frac{(t_{i+1} - t_i)^2}{2}y''(\xi_i),$$

para algún número ξ_i en (t_i, t_{i+1}). Si $h = t_{i+1} - t_i$, entonces

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(\xi_i),$$

y, como $y(t)$ satisface la ecuación diferencial (5.6),

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i). \tag{5.7}$$

El método de Euler construye $w_i \approx y(t_i)$ para cada $i = 1, 2, \ldots, N$, al eliminar el término restante. Por tanto,

$$\begin{aligned} w_0 &= \alpha, \\ w_{i+1} &= w_i + hf(t_i, w_i), \qquad \text{para cada } i = 0, 1, \ldots, N - 1. \end{aligned} \tag{5.8}$$

A la ecuación (5.8) se le llama **ecuación de diferencias** asociada al método de Euler. Como veremos luego en este capítulo, la teoría y la solución de este tipo de ecuaciones nos recuerdan en muchos aspectos a la teoría y la solución de las ecuaciones diferenciales. En el algoritmo 5.1 se pone en ejecución el método de Euler.

ALGORITMO 5.1

Método de Euler

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \qquad a \leq t \leq b, \qquad y(a) = \alpha,$$

en $(N + 1)$ números uniformemente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α.

SALIDA aproximación w a y en los $(N + 1)$ valores de t.

Paso 1 Tome $h = (b - a)/N$;
$t = a$;
$w = \alpha$;
SALIDA (t, w).

Paso 2 Para $i = 1, 2, \ldots, N$ haga pasos 3, 4.

Paso 3 Haga $w = w + hf(t, w)$; *(Calcule w_i.)*
$t = a + ih$. *(Calcule t_i.)*

Paso 4 SALIDA (t, w).

Paso 5 PARAR. ■

Pare interpretar geométricamente el método de Euler, nótese que cuando w_i es una aproximación cercana de $y(t_i)$; la suposición de que el problema está bien planteado implica que

$$f(t_i, w_i) \approx y'(t_i) = f(t_i, y(t_i)).$$

En la figura 5.2(a) aparece la gráfica de la función, donde resalta $y(t_i)$. En la figura 5.2(b) se muestra un paso del método de Euler y en la figura 5.3 una serie de pasos.

EJEMPLO 1 Supongamos que empleamos el método de Euler para aproximar la solución al problema de valor inicial

$$y' = y - t^2 + 1, \qquad 0 \le t \le 2, \qquad y(0) = 0.5,$$

con $N = 10$. Entonces $h = 0.2$, $t_i = 0.2i$, $w_0 = 0.5$, y

$$w_{i+1} = w_i + h(w_i - t_i^2 + 1) = w_i + 0.2[w_i - 0.04i^2 + 1] = 1.2w_i - 0.008i^2 + 0.2,$$

para $i = 0, 1, \ldots, 9$. La solución exacta es $y(t) = (t + 1)^2 - 0.5e^t$. En la tabla 5.1 se muestra la comparación entre los valores aproximados en t_i y los valores reales. ■

Figura 5.2

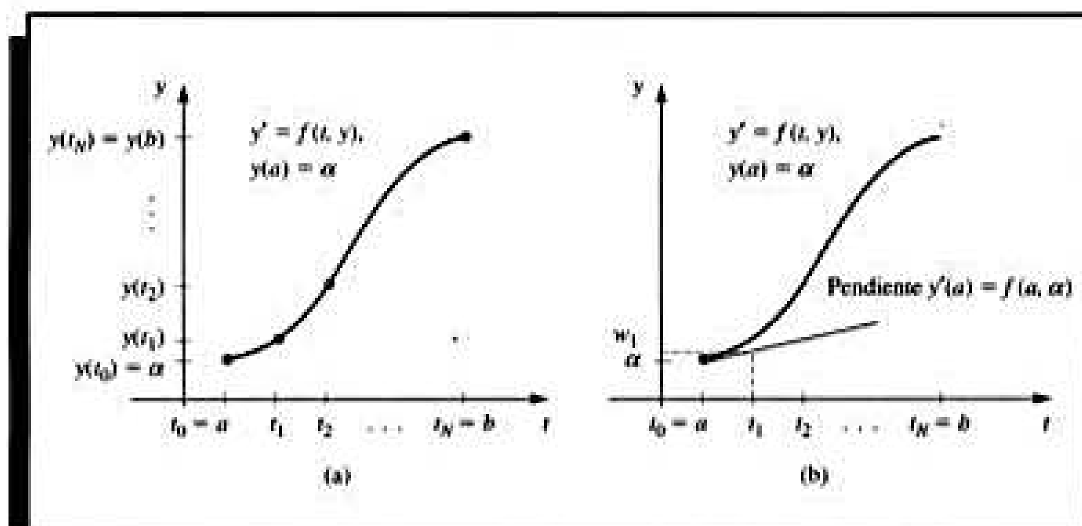


Figura 5.3

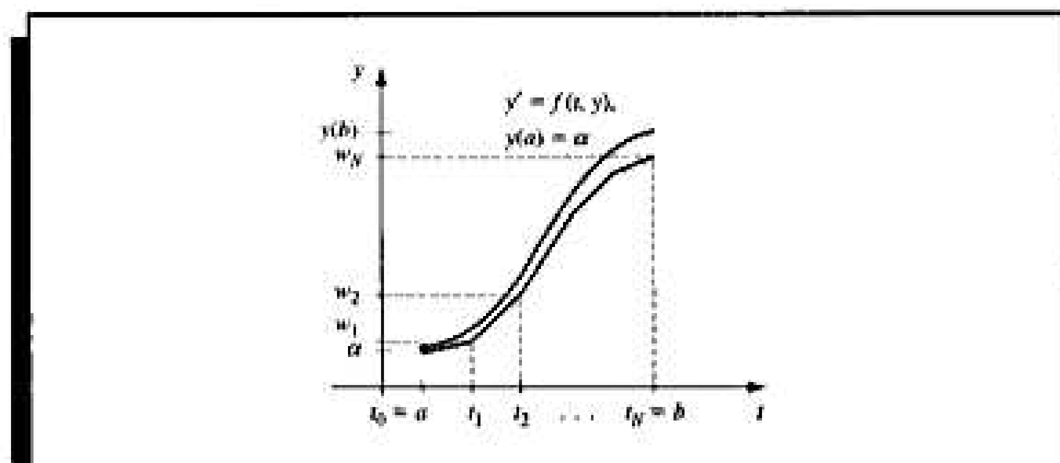


Tabla 5.1

t_i	w_i	$y_i = y(t_i)$	$\lvert y_i - w_i \rvert$
0.0	0.5000000	0.5000000	0.0000000
0.2	0.8000000	0.8292986	0.0292986
0.4	1.1520000	1.2140877	0.0620877
0.6	1.5504000	1.6489406	0.0985406
0.8	1.9884800	2.1272295	0.1387495
1.0	2.4581760	2.6408591	0.1826831
1.2	2.9498112	3.1799415	0.2301303
1.4	3.4517734	3.7324000	0.2806266
1.6	3.9501281	4.2834838	0.3333557
1.8	4.4281538	4.8151763	0.3870225
2.0	4.8657845	5.3054720	0.4396874

Observe que el error crece un poco a medida que el valor de t aumenta. Este crecimiento controlado del error es consecuencia de la estabilidad del método de Euler, el cual implica que se espera que, en el peor de los casos, el error aumente en forma lineal.

Aunque el método de Euler no es lo suficientemente exacto para justificar su uso en la práctica, resulta lo bastante simple para analizar el error producido en su aplicación. El análisis del error con los métodos más precisos que veremos en secciones posteriores sigue el mismo patrón, sólo que es más complicado.

Si queremos obtener una cota de error en el método de Euler, primero consideraremos dos lemas de cálculo.

Lema 5.7 Para toda $x \geq -1$ y para cualquier m positiva, tenemos $0 \leq (1 + x)^m \leq e^{mx}$. ■

Demostración Al aplicar el teorema de Taylor con $f(x) = e^x$, $x_0 = 0$ y $n = 1$ obtenemos

$$e^x = 1 + x + \frac{1}{2}x^2 e^{\xi},$$

donde ξ se encuentra entre x y cero. Por tanto,

$$0 \leq 1 + x \leq 1 + x + \frac{1}{2}x^2e^{\xi} = e^x,$$

y como $1 + x \geq 0$,

$$0 \leq (1 + x)^m \leq (e^x)^m = e^{mx}.$$

■ ■ ■

Lema 5.8 Si s y t son números reales positivos, $\{a_i\}_{i=0}^{k}$ es una sucesión que satisface $a_0 \geq -t/s$, y

$$a_{i+1} \leq (1 + s)a_i + t, \text{ para cada } i = 0, 1, 2, \ldots, k, \tag{5.9}$$

entonces

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

■

Demostración Para un entero fijo i, la desigualdad (5.9) implica que

$$\begin{aligned} a_{i+1} &\leq (1 + s)a_i + t \\ &\leq (1 + s)[(1 + s)a_{i-1} + t] + t \\ &\leq (1 + s)\{(1 + s)[(1 + s)a_{i-2} + t] + t\} + t \\ &\vdots \\ &\leq (1 + s)^{i+1}a_0 + [1 + (1 + s) + (1 + s)^2 + \cdots + (1 + s)^i]\, t. \end{aligned}$$

Pero

$$1 + (1 + s) + (1 + s)^2 + \cdots + (1 + s)^i = \sum_{j=0}^{i}(1 + s)^j$$

es una serie geométrica con razón $(1 + s)$ y, por tanto, su suma es

$$\frac{1 - (1 + s)^{i+1}}{1 - (1 + s)} = \frac{1}{s}[(1 + s)^{i+1} - 1].$$

Por tanto,

$$a_{i+1} \leq (1 + s)^{i+1}a_0 + \frac{(1 + s)^{i+1} - 1}{s}t = (1 + s)^{i+1}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s},$$

y, de acuerdo con el lema 5.7, con $x = 1 + s$ dada

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

■ ■ ■

Teorema 5.9 Supongamos que f es continua y que satisface la condición de Lipschitz con la constante L en

$$D = \{(t, y) \,|\, a \leq t \leq b, -\infty < y < \infty\}$$

y que existe una constante M con la propiedad de que

$$|y''(t)| \leq M, \quad \text{para toda } t \in [a, b].$$

Denotemos con $y(t)$ la solución única del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha$$

y sean $w_0, w_1, \ldots, w_N$ las aproximaciones generadas con el método de Euler para algún entero positivo N. Entonces para cada $i = 0, 1, 2, \ldots, N$,

$$|y(t_i) - w_i| \leq \frac{hM}{2L}[e^{L(t_i-a)} - 1]. \tag{5.10}$$

■

Demostración Cuando $i = 0$, el resultado es verdadero porque $y(t_0) = w_0 = \alpha$.

Conforme a la ecuación (5.7), para $i = 0, 1, \ldots, N-1$ tenemos

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i),$$

y conforme a las ecuaciones en (5.8),

$$w_{i+1} = w_i + hf(t_i, w_i).$$

En consecuencia, al utilizar la notación $y_i = y(t_i)$ y $y_{i+1} = y(t_{i+1})$, tenemos

$$y_{i+1} - w_{i+1} = y_i - w_i + h[f(t_i, y_i) - f(t_i, w_i)] + \frac{h^2}{2}y''(\xi_i)$$

y

$$|y_{i+1} - w_{i+1}| \leq |y_i - w_i| + h|f(t_i, y_i) - f(t_i, w_i)| + \frac{h^2}{2}|y''(\xi_i)|.$$

Puesto que f satisface una condición de Lipschitz en la segunda variable con la constante L y como $|y''(t)| \leq M$, tenemos

$$|y_{i+1} - w_{i+1}| \leq (1 + hL)|y_i - w_i| + \frac{h^2M}{2}.$$

Al referirnos al lema 5.8 y al suponer que $a_j = |y_j - w_j|$ para cada $j = 0, 1, \ldots, N$, y que $s = hL$ y $t = h^2M/2$, tenemos

$$|y_{i+1} - w_{i+1}| \leq e^{(i+1)hL}\left(|y_0 - w_0| + \frac{h^2M}{2hL}\right) - \frac{h^2M}{2hL}.$$

Puesto que $|y_0 - w_0| = 0$ e $(i+1)h = t_{i+1} - t_0 = t_{i+1} - a$, tenemos

$$|y_{i+1} - w_{i+1}| \leq \frac{hM}{2L}(e^{(t_{i+1}-a)L} - 1),$$

para cada $i = 0, 1, \ldots, N-1$. ■ ■ ■

El punto débil del teorema 5.9 consiste en el requisito de conocer una cota de la segunda derivada de la solución. Aunque con frecuencia esta condición nos impide obtener

una cota de error realista, conviene señalar que, si existen $\partial f / \partial t$ y $\partial f / \partial y$, la regla de la cadena para la diferenciación parcial implica que

$$y''(t) = \frac{dy'}{dt}(t) = \frac{df}{dt}(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \cdot f(t, y(t)).$$

Así pues, a veces es posible obtener una cota de error para $y''(t)$ sin que se conozca explícitamente $y(t)$.

EJEMPLO 2 Volviendo ahora al problema de valor inicial

$$y' = y - t^2 + 1, \qquad 0 \le t \le 2, \qquad y(0) = 0.5,$$

considerado en el ejemplo 1, vemos que como $f(t, y) = y - t^2 + 1$, tenemos $\partial f(t, y)/\partial y = 1$ para toda y y, por lo mismo, $L = 1$. En este problema la solución exacta es $y(t) = (t + 1)^2 - \frac{1}{2}e^t$ de manera que $y''(t) = 2 - 0.5e^t$ y

$$|y''(t)| \le 0.5e^2 - 2, \qquad \text{para toda } t \in [0, 2].$$

Al utilizar la desigualdad de la cota de error en el método de Euler con $h = 0.2$, $L = 1$ y $M = 0.5e^2 - 2$ obtenemos la cota de error

$$|y_i - w_i| \le 0.1(0.5e^2 - 2)(e^{t_i} - 1).$$

La tabla 5.2 contiene el error real encontrado en el ejemplo 1, junto con esta cota de error. Nótese que, aunque utilizamos la cota verdadera en la segunda derivada de la solución, la cota de error es mucho mayor que el error real. ■

Tabla 5.2

t_i	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
Error real	0.02930	0.06209	0.09854	0.13875	0.18268	0.23013	0.28063	0.33336	0.38702	0.43969
Cota de error	0.03752	0.08334	0.13931	0.20767	0.29117	0.39315	0.51771	0.66985	0.85568	1.08264

La importancia principal de la fórmula de la cota de error que se da en el teorema 5.9, radica en que la cota depende linealmente del tamaño de paso h. En consecuencia, cuando el tamaño disminuye, deberá haber mayor exactitud en las aproximaciones.

En el resultado del teorema 5.9 no se tiene en cuenta el efecto que el error de redondeo ejerce sobre la elección del tamaño de paso. Conforme h decrece, se requieren más cálculos y se puede predecir un mayor error de redondeo. Así pues, en la práctica la forma de la ecuación de diferencia

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + hf(t_i, w_i), \qquad \text{para cada } i = 0, 1, \ldots, N - 1,$$

no se utiliza para calcular la aproximación a la solución y_i en un punto de red t_i. En cambio, usamos una ecuación de la forma

$$u_0 = \alpha, + \delta_0,$$
$$u_{i+1} = u_i + hf(t_i, u_i) + \delta_{i+1}, \qquad \text{para cada } i = 0, 1, \ldots, N - 1. \qquad (5.11)$$

donde δ_i denota el error de redondeo asociado a u_i. Al usar métodos semejantes a los usados en la demostración del teorema 5.9, podemos producir una cota de error para las aproximaciones de dígitos finitos a y_i obtenidas con el método de Euler.

Teorema 5.10 Sea $y(t)$ la solución única al problema de valor inicial

$$y' = f(t, y), \qquad a \le t \le b, \qquad y(a) = \alpha \tag{5.12}$$

y sean $u_0, u_1, \ldots, u_N$ las aproximaciones obtenidas mediante (5.11). Si $|\delta_i| < \delta$ para cada $i = 0, 1, \ldots, N$ y las hipótesis del teorema 5.9 son aplicables a (5.12), entonces

$$|y(t_i) - u_i| \le \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)[e^{L(t_i-a)} - 1] + |\delta_0| e^{L(t_i-a)}, \tag{5.13}$$

para cada $i = 0, 1, \ldots, N$. ■

La cota de error (5.13) ya no es lineal en h. De hecho, dado que

$$\lim_{h\to 0}\left(\frac{hM}{2} + \frac{\delta}{h}\right) = \infty,$$

se puede esperar que el error se vuelva grande con valores de h suficientemente pequeños. Podemos utilizar el cálculo y determinar una cota más baja para el tamaño de paso h. Tomar $E(h) = (hM/2) + (\delta/h)$ implica que $E'(h) = (M/2) - (\delta/h^2)$.

Si $h < \sqrt{2\delta/M}$, cuando $E'(h) < 0$ y $E(h)$ es decreciente.

Si $h > \sqrt{2\delta/M}$, cuando $E'(h) > 0$ y $E(h)$ es creciente.

El valor mínimo de $E(h)$ ocurre cuando

$$h = \sqrt{\frac{2\delta}{M}}. \tag{5.14}$$

Cuando reducimos h más allá de este valor, el error de la aproximación tiende a incrementarse. No obstante, normalmente el valor de δ es lo bastante pequeño para que esta cota más baja de h no influya en la operación del método de Euler.

CONJUNTO DE EJERCICIOS 5.2

1. Aplique el método de Euler para aproximar las soluciones de los siguientes problemas de valor inicial.
   a. $y' = te^{3t} - 2y, \quad 0 \le t \le 1, \quad y(0) = 0, \quad \text{con } h = 0.5$
   b. $y' = 1 + (t - y)^2, \quad 2 \le t \le 3, \quad y(2) = 1, \quad \text{con } h = 0.5$
   c. $y' = 1 + y/t, \quad 1 \le t \le 2, \quad y(1) = 2, \quad \text{con } h = 0.25$
   d. $y' = \cos 2t + \operatorname{sen} 3t, \quad 0 \le t \le 1, \quad y(0) = 1, \quad \text{con } h = 0.25$
2. A continuación se dan las soluciones reales de los problemas de valor inicial del ejercicio 1. En cada paso, compare el error real con la cota de error.

a. $y(t) = \frac{1}{5}te^{3t} - \frac{1}{25}e^{3t} + \frac{1}{25}e^{-2t}$ **b.** $y(t) = t + \frac{1}{1-t}$

c. $y(t) = t\ln t + 2t$ **d.** $y(t) = \frac{1}{2}\operatorname{sen} 2t - \frac{1}{3}\cos 3t + \frac{4}{3}$

3. Aplique el método de Euler para aproximar las soluciones de los siguientes problemas de valor inicial.

a. $y' = y/t - (y/t)^2$, $1 \le t \le 2$, $y(1) = 1$, con $h = 0.1$

b. $y' = 1 + y/t + (y/t)^2$, $1 \le t \le 3$, $y(1) = 0$, con $h = 0.2$

c. $y' = -(y + 1)(y + 3)$, $0 \le t \le 2$, $y(0) = -2$, con $h = 0.2$

d. $y' = -5y + 5t^2 + 2t$, $0 \le t \le 1$, $y(0) = \frac{1}{3}$, con $h = 0.1$

4. A continuación se dan las soluciones reales a los problemas de valor inicial del ejercicio 3. Calcule el error en las aproximaciones del ejercicio 3.

a. $y(t) = \frac{t}{1 + \ln t}$ **b.** $y(t) = t\tan(\ln t)$

c. $y(t) = -3 + \frac{2}{1 + e^{-2t}}$ **d.** $y(t) = t^2 + \frac{1}{3}e^{-5t}$

5. Dado el problema de valor inicial

$$y' = \frac{2}{t}y + t^2e^t, \quad 1 \le t \le 2, \quad y(1) = 0,$$

con las soluciones exactas $y(t) = t^2(e^t - e)$:

a. Use el método de Euler con $h = 0.1$ para aproximar la solución y compararla con los valores reales de y.

b. Use las respuestas obtenidas en la parte (a) y la interpolación lineal para aproximar los siguientes valores de y y compárelos con los valores reales.

i. $y(1.04)$ **ii.** $y(1.55)$ **iii.** $y(1.97)$

c. Por medio de la ecuación (5.10), calcule el valor de h necesario para que $|y(t_i) - w_i| \le 0.1$.

6. Dado el problema de valor inicial

$$y' = \frac{1}{t^2} - \frac{y}{t} - y^2, \quad 1 \le t \le 2, \quad y(1) = -1,$$

con la solución exacta $y(t) = -1/t$:

a. Use el método de Euler con $h = 0.05$ para aproximar la solución y compárela con los valores reales de y.

b. Use las respuestas obtenidas en el inciso (a) y la interpolación lineal para aproximar los siguientes valores de y y compárelos con los valores reales.

i. $y(1.052)$ **ii.** $y(1.555)$ **iii.** $y(1.978)$

c. Use la ecuación (5.10) para calcular el valor de h necesario para que $|y(t_i) - w_i| \le 0.05$.

7. Dado el problema de valor inicial

$$y' = -y + t + 1, \quad 0 \le t \le 5, \quad y(0) = 1$$

con la solución exacta $y(t) = e^{-t} + t$.

a. Aproxime $y(5)$ aplicando el método de Euler, con $h = 0.2$, $h = 0.1$ y $h = 0.05$.

b. Determine el valor óptimo de h que debe usarse al calcular $y(5)$, suponiendo que $\delta = 10^{-6}$ y que la ecuación (5.14) es válida.

8. Use los resultados del ejercicio 3 y la interpolación lineal para aproximar los siguientes valores de $y(t)$. Compare las aproximaciones obtenidas con los valores reales obtenidos por medio de las funciones del ejercicio 4.

a. $y(1.25)$ y $y(1.93)$ **b.** $y(2.1)$ y $y(2.75)$

c. $y(1.4)$ y $y(1.93)$ **c.** $y(0.54)$ y $y(0.94)$

9. Sea $E(h) = \frac{hM}{2} + \frac{\delta}{h}$.

a. En el problema de valor inicial

$$y' = -y + 1, \qquad 0 \le t \le 1, \qquad y(0) = 0,$$

calcule el valor de h con el que $E(h)$ se reduce al mínimo. Suponga que $\delta = 5 \times 10^{-(n+1)}$, si está empleando la aritmética de n dígitos en el inciso (c).

b. Para el h óptimo calculado en el inciso (a) determine, con la ecuación (5.13), el error mínimo obtenible.

c. Compare el error real obtenido al utilizar $h = 0.1$ y $h = 0.01$ con el error mínimo del inciso (b). ¿Puede explicar los resultados?

10. Considere el problema de valor inicial

$$y' = -10y, \qquad 0 \le t \le 2, \qquad y(0) = 1,$$

que tiene la solución $y(t) = e^{-10t}$. ¿Qué sucede cuando aplicamos el método de Euler a este problema, con $h = 0.1$? ¿Viola este comportamiento el teorema 5.9?

11. En un libro titulado *Looking at History Through Mathematics*, Rashevsky [Ra, pp. 103-110], se propone un modelo de un problema referente a la aparición de no conformistas de la sociedad. Suponga que una sociedad tiene una población de $x(t)$ individuos en el tiempo t, en años, y que todos los no conformistas que tienen relaciones sexuales con otros no conformistas engendran hijos que también son no conformistas, mientras que una proporción fija r del resto de los hijos también son no conformistas. Si supone que las tasas de natalidad y mortalidad de todos los individuos son las constantes b y d, respectivamente, y si los conformistas y no conformistas tienen relaciones sexuales al azar, el problema se puede expresar mediante las ecuaciones diferenciales

$$\frac{dx(t)}{dt} = (b - d)x(t) \quad \text{y} \quad \frac{dx_n(t)}{dt} = (b - d)x_n(t) + rb(x(t) - x_n(t)),$$

donde $x_n(t)$ denota la cantidad de no conformistas de la población en el tiempo t.

a. Si introducimos la variable $p(t) = x_n(t)/x(t)$ para representar la proporción de no conformistas de la sociedad en el tiempo t, demuestre que estas ecuaciones pueden combinarse y simplificarse en la ecuación diferencial individual

$$\frac{dp(t)}{dt} = rb(1 - p(t)).$$

b. Suponiendo que $p(0) = 0.01$, $b = 0.02$, $d = 0.015$ y $r = 0.1$, aproxime la solución $p(t)$ de $t = 0$ a $t = 50$ cuando el tamaño de paso es $h = 1$ año.

c. Resuelva la ecuación diferencial para $p(t)$ exactamente, y compare su resultado del inciso (b) cuando $t = 50$ con el valor exacto en ese tiempo.

12. En un circuito de voltaje impreso $\mathcal{E}$ que tiene la resistencia R, la inductancia L y la capacitancia C en paralelo, la corriente i satisface la ecuación diferencial

$$\frac{di}{dt} = C\frac{d^2\mathcal{E}}{dt^2} + \frac{1}{R}\frac{d\mathcal{E}}{dt} + \frac{1}{L}\mathcal{E}.$$

Supongamos que $C = 0.3$ faradios, $R = 1.4$ ohms, $L = 1.7$ henrios y que el voltaje está dado por

$$\mathcal{E}(t) = e^{-0.06\pi t}\operatorname{sen}(2t - \pi).$$

Si $i(0) = 0$, calcule la corriente i con los valores $t = 0.1j$, donde $j = 0, 1, \ldots, 100$.

5.3 Métodos de Taylor de orden superior

Los métodos numéricos tienen por objeto producir aproximaciones suficientemente exactas con un mínimo de esfuerzo; por ello, necesitamos un medio que nos permita comparar la eficiencia de diversos métodos de aproximación. El primer instrumento que estudiaremos se llama *error local de truncamiento* del método. En un paso específico, este error mide la cantidad en que la solución exacta de la ecuación diferencial no satisface la ecuación de diferencias con que se obtiene la aproximación.

Definición 5.11 El método de diferencias

$$w_0 = \alpha$$
$$w_{i+1} = w_i + h\phi(t_i, w_i), \qquad \text{para cada } i = 0, 1, \ldots, N-1,$$

tiene un **error local de truncamiento** dado por

$$\tau_{i+1}(h) = \frac{y_{i+1} - (y_i + h\phi(t_i, y_i))}{h} = \frac{y_{i+1} - y_i}{h} - \phi(t_i, y_i),$$

para cada $i = 0, 1, \ldots, N-1$. ■

En el método de Euler, el error local de truncamiento para el i-ésimo paso en el problema

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

es

$$\tau_{i+1}(h) = \frac{y_{i+1} - y_i}{h} - f(t_i, y_i), \qquad \text{para cada } i = 0, 1, \ldots, N-1,$$

donde, como siempre, $y_i = y(t_i)$ denota el valor exacto de la solución en t_i.

Este es un *error local*, porque mide la exactitud del método en un paso determinado, suponiendo que el método fue exacto en el paso anterior. Así pues, se basa en la ecuación diferencial, en el tamaño de paso y en el paso particular de la aproximación.

Al estudiar detenidamente la ecuación (5.7) de la sección anterior, observamos que el método de Euler tiene

$$\tau_{i+1}(h) = \frac{h}{2}y''(\xi_i), \qquad \text{para cada } \xi_i \text{ en } (t_i, t_{i+1}).$$

Cuando se sabe que $y''(t)$ está acotado por una constante M en $[a, b]$, ello implica que

$$|\tau_{i+1}(h)| \leq \frac{h}{2}M,$$

así que el error local de truncamiento en el método de Euler es $O(h)$.

Una manera de seleccionar los métodos de la ecuación de diferencias para resolver ecuaciones diferenciales ordinarias, es hacerlo de manera que sus errores locales de truncamiento sean $O(h^p)$ con el valor de p más grande posible, sin que el número y la complejidad de los cálculos de los métodos rebasen una cota razonable.

Obtenemos el método de Euler aplicando el teorema de Taylor con $n = 1$, para aproximar la solución de la ecuación diferencial; por ello, nuestro primer intento de encontrar métodos que mejoren las propiedades de convergencia de los métodos de diferencia consiste en ampliar esta técnica de derivación para valores mayores de n.

Supongamos que la solución $y(t)$ del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

tiene $(n + 1)$ derivadas continuas. Si se desarrolla la solución $y(t)$ en función del n-ésimo polinomio de Taylor alrededor de t_i y calculamos en t_{i+1} obtendremos

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(t_i) + \cdots + \frac{h^n}{n!}y^{(n)}(t_i) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_i), \tag{5.15}$$

para alguna ξ_i en (t_i, t_{i+1}).

La diferenciación sucesiva de la solución $y(t)$ nos da

$$y'(t) = f(t, y(t)),$$
$$y''(t) = f'(t, y(t)),$$

y, en general,

$$y^{(k)}(t) = f^{(k-1)}(t, y(t)).$$

Al sustituir estos resultados en la ecuación (5.15), obtenemos

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}f'(t_i, y(t_i)) + \cdots \tag{5.16}$$

$$+ \frac{h^n}{n!}f^{(n-1)}(t_i, y(t_i)) + \frac{h^{n+1}}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)).$$

El método de la ecuación de diferencias correspondiente a la ecuación (5.16) recibe el nombre de *método de Taylor de orden n* y se obtiene suprimiendo el término residual que contiene ξ_i.

Método de Taylor de orden n:

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + hT^{(n)}(t_i, w_i) \quad \text{para cada } i = 0, 1, \ldots, N-1. \tag{5.17}$$

donde

$$T^{(n)}(t_i, w_i) = f(t_i, w_i) + \frac{h}{2} f'(t_i, w_i) + \cdots + \frac{h^{n-1}}{n!} f^{(n-1)}(t_i, w_i).$$

Nótese que el método de Euler es el método de Taylor de orden uno.

EJEMPLO 1 Para aplicar el método de Taylor de órdenes dos y cuatro al problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

que estudiamos en las secciones anteriores, debemos encontrar las tres primeras derivadas de $f(t, y(t)) = y(t) - t^2 + 1$ respecto a la variable t:

$$f'(t, y(t)) = \frac{d}{dt}(y - t^2 + 1) = y' - 2t = y - t^2 + 1 - 2t,$$

$$f''(t, y(t)) = \frac{d}{dt}(y - t^2 + 1 - 2t) = y' - 2t - 2$$

$$= y - t^2 + 1 - 2t - 2 = y - t^2 - 2t - 1,$$

y

$$f'''(t, y(t)) = \frac{d}{dt}(y - t^2 - 2t - 1) = y' - 2t - 2 = y - t^2 - 2t - 1.$$

Entonces

$$T^{(2)}(t_i, w_i) = f(t_i, w_i) + \frac{h}{2} f'(t_i, w_i) = w_i - t_i^2 + 1 + \frac{h}{2}(w_i - t_i^2 - 2t_i + 1)$$

$$= \left(1 + \frac{h}{2}\right)(w_i - t_i^2 + 1) - ht_i$$

y

$$T^{(4)}(t_i, w_i) = f(t_i, w_i) + \frac{h}{2} f'(t_i, w_i) + \frac{h^2}{6} f''(t_i, w_i) + \frac{h^3}{24} f'''(t_i, w_i)$$

$$= w_i - t_i^2 + 1 + \frac{h}{2}(w_i - t_i^2 - 2t_i + 1) + \frac{h^2}{6}(w_i - t_i^2 - 2t_i - 1)$$

$$+ \frac{h^3}{24}(w_i - t_i^2 - 2t_i - 1)$$

$$= \left(1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^3}{24}\right)(w_i - t_i^2) - \left(1 + \frac{h}{3} + \frac{h^2}{12}\right)(ht_i)$$

$$+ 1 + \frac{h}{2} - \frac{h^2}{6} - \frac{h^3}{24}.$$

En consecuencia, los métodos de Taylor de órdenes dos y cuatro son

$$w_0 = 0.5,$$

$$w_{i+1} = w_i + h\left[\left(1 + \frac{h}{2}\right)(w_i - t_i^2 + 1) - ht_i\right]$$

y

$$w_0 = 0.5,$$

$$w_{i+1} = w_i + h\left[\left(1 + \frac{h}{2} + \frac{h^2}{6} + \frac{h^3}{24}\right)(w_i - t_i^2) - \left(1 + \frac{h}{3} + \frac{h^2}{12}\right)ht_i + 1 + \frac{h}{2} - \frac{h^2}{6} - \frac{h^3}{24}\right],$$

para cada $i = 0, 1, \ldots, N-1$.

Si $h = 0.2$, entonces $N = 10$ y $t_i = 0.2i$ para $i = 1, 2, \ldots, 10$. Por tanto, el método de segundo orden se convierte en

$$w_0 = 0.5,$$

$$w_{i+1} = w_i + 0.2\left[\left(1 + \frac{0.2}{2}\right)(w_i - 0.04i^2 + 1) - 0.04i\right]$$

$$= 1.22w_i - 0.0088i^2 - 0.008i + 0.22,$$

y el método de cuarto orden se convierte en

$$w_{i+1} = w_i + 0.2\left[\left(1 + \frac{0.2}{2} + \frac{0.04}{6} + \frac{0.008}{24}\right)(w_i - 0.04i^2) - \left(1 + \frac{0.2}{3} + \frac{0.04}{12}\right)(0.04i) + 1 + \frac{0.2}{2} - \frac{0.04}{6} - \frac{0.008}{24}\right]$$

$$= 1.2214w_i - 0.008856i^2 - 0.00856i + 0.2186,$$

para cada $i = 0, 1, \ldots, 9$.

La tabla 5.3 contiene los valores reales de la solución $y(t) = (t+1)^2 - 0.5e^t$, los resultados obtenidos con los métodos de Taylor de orden dos y cuatro y los errores reales a que dan origen esos métodos.

Tabla 5.3

t_i	Valores exactos $y(t_i)$	Métodos de Taylor de orden 2 w_i	Error $\lvert y(t_i) - w_i\rvert$	Métodos de Taylor de orden 4 w_i	Error $\lvert y(t_i) - w_i\rvert$
0.0	0.5000000	0.5000000	0	0.5000000	0
0.2	0.8292986	0.8300000	0.0007014	0.8293000	0.0000014
0.4	1.2140877	1.2158000	0.0017123	1.2140910	0.0000034
0.6	1.6489406	1.6520760	0.0031354	1.6489468	0.0000062
0.8	2.1272295	2.1323347	0.0051032	2.1272396	0.0000101
1.0	2.6408591	2.6486459	0.0077868	2.6408744	0.0000153
1.2	3.1799415	3.1913480	0.0114065	3.1799640	0.0000225
1.4	3.7324000	3.7486446	0.0162446	3.7324321	0.0000321
1.6	4.2834838	4.3061464	0.0226626	4.2835285	0.0000447
1.8	4.8151763	4.8462986	0.0311223	4.8152377	0.0000615
2.0	5.3054720	5.3476843	0.0422123	5.3055554	0.0000834

Supóngase que debemos encontrar una aproximación a un punto intermedio de la tabla, por ejemplo en $t = 1.25$. Si empleamos la interpolación lineal en las aproximaciones mediante el método de Taylor de orden cuatro en $t = 1.2$ y en $t = 1.4$, tenemos

$$y(1.25) \approx \left(\frac{1.25 - 1.4}{1.2 - 1.4}\right)3.1799640 + \left(\frac{1.25 - 1.2}{1.4 - 1.2}\right)3.7324321 = 3.3180810.$$

Puesto que $y(1.25) = 3.3173285$, esta aproximación tiene un error de 0.0007525, cifra que es casi 30 veces el promedio de los errores de aproximación en 1.2 y en 1.4.

Si queremos mejorar la aproximación a $y(1.25)$, podemos usar la interpolación cúbica de Hermite. Para ello se requieren aproximaciones a $y'(1.2)$ y $y'(1.4)$ y también aproximaciones a $y(1.2)$ y $y(1.4)$. Pero las aproximaciones de la derivada están disponibles de la ecuación diferencial, porque $y'(t) = f(t, y(t))$. En nuestro ejemplo, ello significa que $y'(t) = y(t) - t^2 + 1$, de modo que

$$y'(1.2) = y(1.2 - (1.2)^2 + 1 \approx 3.1799640 - 1.44 + 1 = 2.7399640$$

y

$$y'(1.4) = y(1.4) - (1.4)^2 + 1 \approx 3.7324327 - 1.96 + 1 = 2.7724321.$$

Los resultados que se muestran en la tabla 5.4 fueron obtenidos con el procedimiento de las diferencias divididas de la sección 3.3. Las entradas subrayadas provienen de los datos, las entradas restantes son resultado del uso de las fórmulas del procedimiento de las diferencias divididas.

Tabla 5.4

1.2	<u>3.1799640</u>			
		<u>2.7399640</u>		
1.2	<u>3.1799640</u>		0.1118825	
		2.7623405		−0.3071225
1.4	<u>3.7324321</u>		0.0504580	
		<u>2.7724321</u>		
1.4	<u>3.7324321</u>			

El polinomio cúbico de Hermite es

$$y(t) \approx 3.1799640 + (t - 1.2)2.7399640 + (t - 1.2)^2 0.1118825 \\ + (t - 1.2)^2(t - 1.4)(-0.3071225),$$

y, por tanto,

$$y(1.25) \approx 3.1799640 + 0.1369982 + 0.0002797 + 0.0001152 = 3.3173571,$$

resultado que tiene una exactitud de 0.0000286. Esto es más o menos el promedio del error en 1.2 y en 1.4, que está cerca de 4% del error obtenido cuando se usa la interpolación lineal. ■

Teorema 5.12 Si se utiliza el método de Taylor de orden n para aproximar la solución de

$$y'(t) = f(t, y(t)), \qquad a \leq t \leq b, \qquad y(a) = \alpha,$$

con tamaño de paso h y si $y \in C^{n+1}[a, b]$, entonces el error local de truncamiento es $O(h^n)$. ■

Demostración Note que la ecuación 5.16 puede rescribirse como

$$y_{i+1} - y_i - hf(t_i, y_i) - \frac{h^2}{2} f'(t_i, y_i) - \cdots - \frac{h^n}{n!} f^{(n-1)}(t_i, y_i) = \frac{h^{n+1}}{(n+1)!} f^{(n)}(\xi_i, y(\xi_i)),$$

para alguna ξ_i en (t_i, t_{i+1}). Así que el error local de truncamiento es

$$\tau_{i+1}(h) = \frac{y_{i+1} - y_i}{h} - T^{(n)}(t_i, y_i) = \frac{h^n}{(n+1)!} f^{(n)}(\xi_i, y(\xi_i)),$$

para cada $i = 0, 1, \ldots, N-1$. Si $y \in C^{n+1}[a, b]$, ello significa que $y^{(n+1)}(t) = f^{(n)}(t, y(t))$ está acotado en $[a, b]$ y que $\tau_i = O(h^n)$ para cada $i = 1, 2, \ldots, N$. ■ ■ ■

CONJUNTO DE EJERCICIOS 5.3

1. Aplique el método de Taylor de orden dos para aproximar las soluciones en los siguientes problemas de valor inicial.
   a. $y' = te^{3t} - 2y, \quad 0 \le t \le 1, \quad y(0) = 0$, con $h = 0.5$
   b. $y' = 1 + (t - y)^2, \quad 2 \le t \le 3, \quad y(2) = 1$, con $h = 0.5$
   c. $y' = 1 + y/t, \quad 1 \le t \le 2, \quad y(1) = 2$, con $h = 0.25$
   d. $y' = \cos 2t + \text{sen}\, 3t, \quad 0 \le t \le 1, \quad y(0) = 1$, con $h = 0.25$
2. Repita el ejercicio 1 usando el método de Taylor de orden cuatro.
3. Aplique el método de Taylor de órdenes dos y cuatro para aproximar la solución de los siguientes problemas de valor inicial.
   a. $y' = y/t - (y/t)^2, \quad 1 \le t \le 1.2, \quad y(1) = 1$, con $h = 0.1$
   b. $y' = \text{sen}\, t + e^{-t}, \quad 0 \le t \le 1, \quad y(0) = 0$, con $h = 0.5$
   c. $y' = 1/t(y^2 + y), \quad 1 \le t \le 3, \quad y(1) = -2$, con $h = 0.5$
   d. $y' = -ty + 4t/y, \quad 0 \le t \le 1, \quad y(0) = 1$, con $h = 0.25$
4. Aplique el método de Taylor de orden dos con $h = 0.1$ para aproximar la solución de

$$y' = 1 + t\,\text{sen}(ty), \quad 0 \le t \le 2, \quad y(0) = 0.$$

5. Dado el problema de valor inicial

$$y' = \frac{2}{t} y + t^2 e^t, \qquad 1 \le t \le 2, \qquad y(1) = 0,$$

   con la solución exacta $y(t) = t^2(e^t - e)$:
   a. Aplique el método de Taylor de orden dos con $h = 0.1$ para aproximar la solución y compárela con los valores reales de y.
   b. Use las respuestas obtenidas en el inciso (a) y la interpolación lineal para aproximar y en los siguientes valores y compárelos con los valores reales de y.
      i. $y(1.04)$ ii. $y(1.55)$ iii. $y(1.97)$

c. Aplique el método de Taylor de orden cuatro con $h = 0.1$ para aproximar la solución y compárela con los valores reales de y.

d. Use las respuestas obtenidas en el inciso (c) y la interpolación cúbica fragmentaria de Hermite para aproximar y en los siguientes valores y compárelos con los valores reales de y.

i. $y(1.04)$ ii. $y(1.55)$ iii. $y(1.97)$

6. Dado el problema de valor inicial

$$y' = \frac{1}{t^2} - \frac{y}{t} - y^2, \qquad 1 \le t \le 2, \qquad y(1) = -1,$$

con la solución exacta $y(t) = -1/t$:

a. Aplique el método de Taylor de orden dos con $h = 0.05$ para aproximar la solución y compárela con los valores reales de y.

b. Use las respuestas obtenidas en el inciso (a) y la interpolación lineal para aproximar los siguientes valores de y, y después compárelos con los valores reales.

i. $y(1.052)$ ii. $y(1.555)$ iii. $y(1.978)$

c. Aplique el método de Taylor de orden cuatro con $h = 0.05$ para aproximar la solución y compárela con los valores reales de y.

d. Use las respuestas generadas en el inciso (c) y la interpolación cúbica de Hermite para aproximar los siguientes valores de y, y compárelos con los valores reales.

i. $y(1.052)$ ii. $y(1.555)$ iii. $y(1.978)$

7. Un proyectil de masa $m = 0.11$ kg que es lanzado verticalmente hacia arriba con una velocidad inicial $v(0) = 8$ m/s, disminuye su velocidad por efecto de la fuerza de gravedad $F_g = -mg$ y por la resistencia del aire $F_r = kv|v|$, donde $g = 9.8$ m/s² y $k = 0.002$ kg/m. La ecuación diferencial de la velocidad v está dada por

$$mv' = -mg - kv|v|.$$

a. Calcule la velocidad después de 0.1, 0.2, …, 1.0 s.

b. Determine, con una precisión de décimas de segundo, cuándo alcanzará el proyectil su altura máxima y cuándo empezará a caer.

5.4 Métodos de Runge-Kutta

Los métodos de Taylor que vimos en la sección anterior tienen un error local de truncamiento de orden alto, pero poseen la desventaja de requerir el cálculo y evaluación de las derivadas de $f(t, y)$. Este es un procedimiento lento y complicado en la mayor parte de los problemas, por lo cual los métodos de Taylor rara vez se emplean en la práctica.

Los **métodos de Runge-Kutta** tienen el error local de truncamiento de orden alto, como los métodos de Taylor, pero permiten prescindir del cálculo y evaluación de las derivadas de $f(t, y)$. Antes de exponer las ideas en que se funda su deducción debemos enunciar el teorema de Taylor para dos variables. La demostración de este resultado viene en cualquier libro de cálculo avanzado (véase, por ejemplo, a [Fu, p. 331]).

Teorema 5.13 Supóngase que $f(t, y)$ y todas sus derivadas parciales de orden menor o igual que $n + 1$ son continuas en $D = \{(t, y) \mid a \le t \le b, c \le y \le d\}$, y sea $(t_0, y_0) \in D$. Para toda $(t, y) \in D$, existen ξ entre t y t_0 y μ entre y y y_0 con:

$$f(t, y) = P_n(t, y) + R_n(t, y),$$

donde

$$\begin{aligned}P_n(t, y) = f(t_0, y_0) &+ \left[(t - t_0)\frac{\partial f}{\partial t}(t_0, y_0) + (y - y_0)\frac{\partial f}{\partial y}(t_0, y_0)\right] \\ &+ \left[\frac{(t - t_0)^2}{2}\frac{\partial^2 f}{\partial t^2}(t_0, y_0) + (t - t_0)(y - y_0)\frac{\partial^2 f}{\partial t\,\partial y}(t_0, y_0)\right. \\ &+ \left.\frac{(y - y_0)^2}{2}\frac{\partial^2 f}{\partial y^2}(t_0, y_0)\right] + \cdots \\ &+ \left[\frac{1}{n!}\sum_{j=0}^{n}\binom{n}{j}(t - t_0)^{n-j}(y - y_0)^j\frac{\partial^n f}{\partial t^{n-j}\,\partial y^j}(t_0, y_0)\right]\end{aligned}$$

y

$$R_n(t, y) = \frac{1}{(n+1)!}\sum_{j=0}^{n+1}\binom{n+1}{j}(t - t_0)^{n+1-j}(y - y_0)^j\frac{\partial^{n+1} f}{\partial t^{n+1-j}\,\partial y^j}(\xi, \mu).$$

■

A la función $P_n(t, y)$ se le llama **n-ésimo polinomio de Taylor en dos variables** para la función f alrededor de (t_0, y_0) y $R_n(t, y)$ es el término residual asociado a $P_n(t, y)$.

EJEMPLO 1 En la figura 5.4 se muestra la gráfica de la función

$$f(t, y) = \exp\left[-\frac{(t-2)^2}{4} - \frac{(y-3)^2}{4}\right]\cos(2t + y - 7)$$

junto con el segundo polinomio de Taylor de f alrededor de (2, 3), es decir, el polinomio en dos variables

$$P_2(t, y) = 1 - \frac{9}{4}(t-2)^2 - 2(t-2)(y-3) - \frac{3}{4}(y-3)^2.$$

Sería tedioso obtener manualmente la diferenciación necesaria para determinar este polinomio. Por fortuna, disponemos de un procedimiento de Maple que lo hace por nosotros. Primero debemos iniciar el procedimiento del polinomio de Taylor de variables múltiples introduciendo el comando:

```
> readlib (mtaylor);
```

que produce la respuesta

```
proc()... fin proc
```

El polinomio de Taylor que necesitamos en este ejemplo lo obtenemos introduciendo el comando

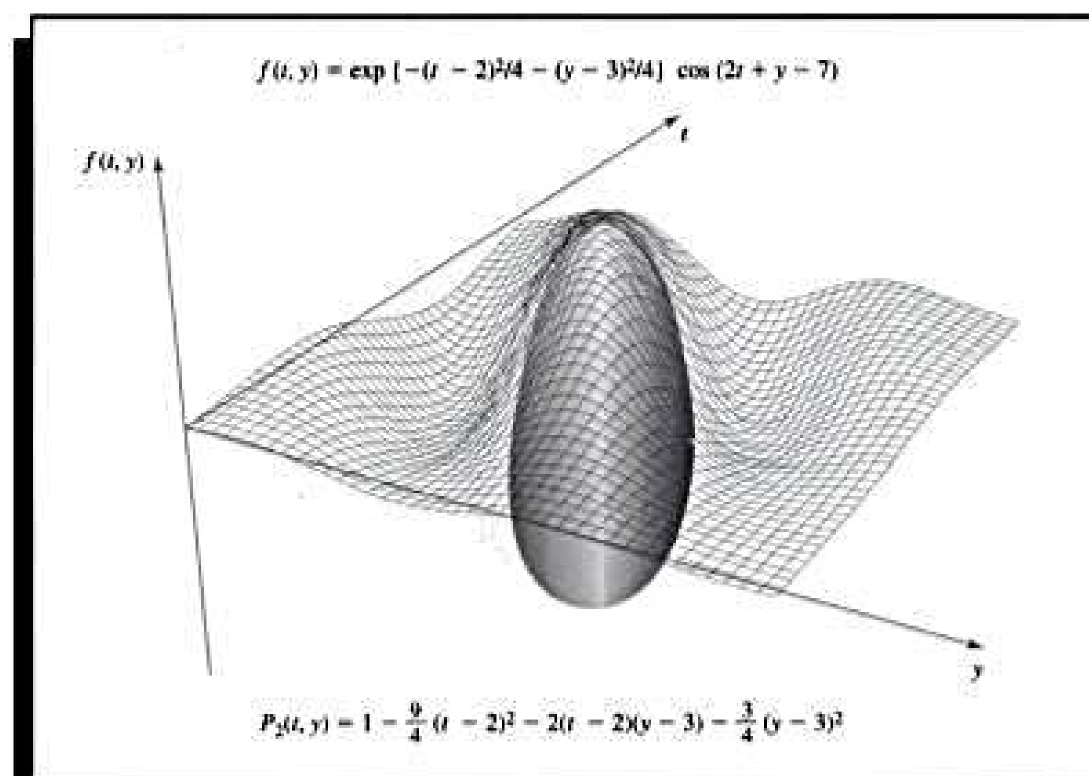
```
>mtaylor(exp(-(t-2)^2/4-(y-3)^2/4)*cos(2*t+y-7),[t=2,y=3],3);
```

El último parámetro de este comando indica que queremos el segundo polinomio multivariado de Taylor, es decir, el polinomio cuadrático. Si este parámetro es 2, obtenemos el polinomio lineal; si es 1, obtenemos el polinomio constante. Cuando se omite este parámetro, adquiere el valor por omisión de 6 y nos da el quinto polinomio.

La respuesta que se obtiene con este comando de Maple es el polinomio

$$1 - \frac{9}{4}(t-2)^2 - 2(t-2)(y-3) - \frac{3}{4}(y-3)^2. \qquad \blacksquare$$

Figura 5.4



El primer paso al derivar el método de Runge-Kutta, es determinar los valores de a_1, α_1 y β_1 con la propiedad de que $a_1 f(t + \alpha_1, y + \beta_1)$ aproxima

$$T^{(2)}(t, y) = f(t, y) + \frac{h}{2} f'(t, y),$$

con un error no mayor que $O(h^2)$, o sea el error local de truncamiento del método de Taylor de orden dos. Dado que

$$f'(t, y) = \frac{df}{dt}(t, y) = \frac{\partial f}{\partial t}(t, y) + \frac{\partial f}{\partial y}(t, y) \cdot y'(t) \quad \text{y} \quad y'(t) = f(t, y),$$

esto implica que

$$T^{(2)}(t, y) = f(t, y) + \frac{h}{2}\frac{\partial f}{\partial t}(t, y) + \frac{h}{2}\frac{\partial f}{\partial y}(t, y) \cdot f(t, y). \tag{5.18}$$

Al desarrollar $f(t + \alpha_1, y + \beta_1)$ en su polinomio de Taylor de grado uno alrededor de (t, y), se obtiene

$$a_1 f(t + \alpha_1, y + \beta_1) = a_1 f(t, y) + a_1\alpha_1 \frac{\partial f}{\partial t}(t, y) + a_1\beta_1 \frac{\partial f}{\partial y}(t, y) + a_1 \cdot R_1(t + \alpha_1, y + \beta_1), \tag{5.19}$$

donde

$$R_1(t + \alpha_1, y + \beta_1) = \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\xi, \mu) + \alpha_1\beta_1 \frac{\partial^2 f}{\partial t\,\partial y}(\xi, \mu) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial y^2}(\xi, \mu), \tag{5.20}$$

para alguna ξ entre t y $t + \alpha_1$ y μ entre y y $y + \beta_1$.

Al igualar los coeficientes de f y sus derivadas en las ecuaciones (5.18) y (5.19), obtenemos las tres ecuaciones

$$f(t, y): \quad a_1 = 1; \qquad \frac{\partial f}{\partial t}(t, y): \quad a_1\alpha_1 = \frac{h}{2};$$

y

$$\frac{\partial f}{\partial y}(t, y): \quad a_1\beta_1 = \frac{h}{2} f(t, y).$$

En forma única se determina que los parámetros a_1, α_1 y β_1 son

$$a_1 = 1, \qquad \alpha_1 = \frac{h}{2} \quad \text{y} \quad \beta_1 = \frac{h}{2} f(t, y);$$

por tanto

$$T^{(2)}(t, y) = f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right) - R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right),$$

y de acuerdo con la ecuación (5.20),

$$R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right) = \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi, \mu) + \frac{h^2}{4} f(t, y)\frac{\partial^2 f}{\partial t\,\partial y}(\xi, \mu) + \frac{h^2}{8}(f(t, y))^2 \frac{\partial^2 f}{\partial y^2}(\xi, \mu).$$

Si todas las derivadas parciales de segundo orden de f están acotadas, entonces

$$R_1\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right)$$

es $O(h^2)$, o sea el orden del error local de truncamiento del método de Taylor de orden dos. En consecuencia, al utilizar el procedimiento nuevo en vez del método de Taylor de orden dos se podría agregar algún error; pero ello no aumenta el orden del error.

El método de la ecuación de diferencia que resulta al sustituir $T^{(2)}(t, y)$ por $f(t + (h/2), y + (h/2)f(t, y))$ en el método de Taylor de orden dos es un método específico de Runge-Kutta, conocido con el nombre de *método del punto medio*.

Método del punto medio:

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + hf\left(t_i + \frac{h}{2}, w_i + \frac{h}{2} f(t_i, w_i)\right), \qquad \text{para cada } i = 0, 1, \ldots, N-1.$$

Puesto que sólo tres parámetros se encuentran en $a_1 f(t + \alpha_1, y + \beta_1)$ y los tres se requieren en la igualdad con $T^{(2)}$, necesitamos una forma más compleja para cumplir las condiciones que requiere cualquiera de los métodos de Taylor de orden superior.

La forma más apropiada de cuatro parámetros con que se aproxima

$$T^{(3)}(t, y) = f(t, y) + \frac{h}{2} f'(t, y) + \frac{h^2}{6} f''(t, y)$$

es

$$a_1 f(t, y) + a_2 f(t + \alpha_2, y + \delta_2 f(t, y)); \tag{5.21}$$

y ni siquiera con esto se tiene la suficiente flexibilidad para igualar el término

$$\frac{h^2}{6}\left[\frac{\partial f}{\partial y}(t, y)\right]^2 f(t, y),$$

resultante de la expansión de $(h^2/6)f''(t, y)$. En consecuencia, lo mejor que podemos lograr utilizando (5.21), son métodos con el error local de truncamiento $O(h^2)$. No obstante, el hecho de que (5.21) tenga cuatro parámetros, da cierta flexibilidad en su elección para poder derivar varios métodos $O(h^2)$. Uno de los más importantes es el *método modificado de Euler*, que corresponde a seleccionar $a_1 = a_2 = \frac{1}{2}$ y $\alpha_2 = \delta_2 = h$ y presenta la siguiente forma de ecuación de diferencias.

Método modificado de Euler:

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + \frac{h}{2}[f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))],$$
$$\text{para cada} \quad i = 0, 1, 2, \ldots, N-1.$$

El otro método importante $O(h^2)$ es el *de Heun*, que corresponde a $a_1 = \frac{1}{4}$, $a_2 = \frac{3}{4}$ y $\alpha_2 = \delta_2 = \frac{2}{3}h$, y que tiene la forma de ecuación de diferencia siguiente.

Método de Heun:

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + \frac{h}{4}[f(t_i, w_i) + 3f(t_i + \frac{2}{3}h, w_i + \frac{2}{3} hf(t_i, w_i))],$$
$$\text{para cada } i = 0, 1, 2, \ldots, N-1.$$

Ambos se clasifican como métodos de Runge-Kutta de orden dos, que es el orden de su error local de truncamiento.

EJEMPLO 2 Supóngase que aplicamos los métodos de Runge-Kutta de orden dos a nuestro ejemplo

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

con $N = 10$, $h = 0.2$, $t_i = 0.2i$ y $w_0 = 0.5$ en cada caso. Las ecuaciones de diferencias son

$$\text{Método del punto medio: } w_{i+1} = 1.22w_i - 0.0088i^2 - 0.008i + 0.218;$$

$$\text{Método modificado de Euler: } w_{i+1} = 1.22w_i - 0.0088i^2 - 0.008i + 0.216;$$

$$\text{Método de Heun: } w_{i+1} = 1.22w_i - 0.0088i^2 - 0.008i + 0.217\bar{3},$$

para cada $i = 0, 1, \ldots, 9$. La tabla 5.5 contiene los resultados de estos cálculos. Para este problema, es mejor el método del punto medio, seguido por el método de Heun. ∎

Tabla 5.5

t_i	$y(t_i)$	Método del punto medio	Error	Método modificado de Euler	Error	Método de Heun	Error
0.0	0.5000000	0.5000000	0	0.5000000	0	0.5000000	0
0.2	0.8292986	0.8280000	0.0012986	0.8260000	0.0032986	0.8273333	0.0019653
0.4	1.2140877	1.2113600	0.0027277	1.2069200	0.0071677	1.2098800	0.0042077
0.6	1.6489406	1.6446592	0.0042814	1.6372424	0.0116982	1.6421869	0.0067537
0.8	2.1272295	2.1212842	0.0059453	2.1102357	0.0169938	2.1176014	0.0096281
1.0	2.6408591	2.6331668	0.0076923	2.6176876	0.0231715	2.6280070	0.0128521
1.2	3.1799415	3.1704634	0.0094781	3.1495789	0.0303627	3.1635019	0.0164396
1.4	3.7324000	3.7211654	0.0112346	3.6936862	0.0387138	3.7120057	0.0203944
1.6	4.2834838	4.2706218	0.0128620	4.2350972	0.0483866	4.2587802	0.0247035
1.8	4.8151763	4.8009586	0.0142177	4.7556185	0.0595577	4.7858452	0.0293310
2.0	5.3054720	5.2903695	0.0151025	5.2330546	0.0724173	5.2712645	0.0342074

Aunque podemos aproximar $T^{(3)}(t, y)$ con el error $O(h^3)$ mediante una expresión de la forma

$$f(t + \alpha_1, y + \delta_1 f(t + \alpha_2, y + \delta_2 f(t, y))),$$

que contiene cuatro parámetros, el álgebra con que se determinan α_1, δ_1, α_2 y δ_2 es complicada y, por lo mismo, no la explicaremos aquí. De hecho, el método de Runge-Kutta de orden tres que resulta de esta expresión generalmente no se emplea. El método de Runge-Kutta de mayor uso es el de orden cuatro y, en la forma de la ecuación en diferencias, se da por el siguiente método.

Método de Runge-Kutta de orden cuatro:

$$w_0 = \alpha,$$

$$k_1 = hf(t_i, w_i),$$

$$k_2 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1\right),$$

$$k_3 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right),$$
$$k_4 = hf(t_{i+1}, w_i + k_3),$$
$$w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

para cada $i = 0, 1, \ldots, N-1$. Este método tiene el error local de truncamiento $O(h^4)$, siempre que la solución $y(t)$ tenga cinco derivadas continuas. Se introduce la notación k_1, k_2, k_3, k_4 en él para prescindir de las anidaciones sucesivas en la segunda variable de $f(t, y)$ (véase el ejercicio 17). En el algoritmo 5.2 se pone en ejecución el método de Runge-Kutta de orden cuatro.

ALGORITMO 5.2

Método de Runge-Kutta de orden cuatro

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

en $(N + 1)$ números uniformemente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α.

SALIDA aproximación w a y en los $(N + 1)$ valores de t.

Paso 1 Tome $h = (b - a)/N$;
$t = a$;
$w = \alpha$;
SALIDA (t, w).

Paso 2 Para $i = 1, 2, \ldots, N$ haga pasos 3-5.

Paso 3 Tome $K_1 = hf(t, w)$;
$K_2 = hf(t + h/2, w + K_1/2)$;
$K_3 = hf(t + h/2, w + K_2/2)$;
$K_4 = hf(t + h, w + K_3)$.

Paso 4 Tome $w = w + (K_1 + 2K_2 + 2K_3 + K_4)/6$; (*Calcule* w_i.)
$t = a + ih$. (*Calcule* t_i.)

Paso 5 SALIDA (t, w).

Paso 6 PARAR. ■

EJEMPLO 3 Al aplicar el método de Runge-Kutta de orden cuatro para obtener aproximaciones a la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

con $h = 0.2$, $N = 10$ y $t_i = 0.2i$ obtenemos los resultados y los errores que se proporcionan en la tabla 5.6. ■

Tabla 5.6

t_i	Valores exactos $y_i = y(t_i)$	Método de Runge-Kutta de orden cuarto w_i	Error $\|y_i - w_i\|$
0.0	0.5000000	0.5000000	0
0.2	0.8292986	0.8292933	0.0000053
0.4	1.2140877	1.2140762	0.0000114
0.6	1.6489406	1.6489220	0.0000186
0.8	2.1272295	2.1272027	0.0000269
1.0	2.6408591	2.6408227	0.0000364
1.2	3.1799415	3.1798942	0.0000474
1.4	3.7324000	3.7323401	0.0000599
1.6	4.2834838	4.2834095	0.0000743
1.8	4.8151763	4.8150857	0.0000906
2.0	5.3054720	5.3053630	0.0001089

El mayor esfuerzo de cálculo que se requiere para aplicar los métodos de Runge-Kutta es la evaluación de f. En los métodos de segundo orden, el error local de truncamiento es $O(h^2)$, y el costo es realizar dos evaluaciones funcionales por paso. El método de Runge-Kutta de orden cuatro requiere cuatro evaluaciones por paso y el error local de truncamiento es $O(h^4)$. Butcher (véase un resumen en [But]) estableció la relación entre la cantidad de evaluaciones por paso y el orden del error local de truncamiento que aparece en la tabla 5.7. En ésta se indica por qué los métodos de un orden menor que cinco con un tamaño menor de paso se prefieren a los de orden superior con un tamaño mayor de paso.

Tabla 5.7

Evaluaciones por paso	2	3	4	$5 \leq n \leq 7$	$8 \leq n \leq 9$	$10 \leq n$
El mejor error local de truncamiento posible	$O(h^2)$	$O(h^3)$	$O(h^4)$	$O(h^{n-1})$	$O(h^{n-2})$	$O(h^{n-3})$

Una medida con que se comparan los métodos de orden menor de Runge-Kutta se describe así:

> Como el método de Runge-Kutta de orden cuatro requiere realizar cuatro evaluaciones por paso, deberá dar respuestas más exactas que las del método de Euler con un cuarto del tamaño de paso para que sea mejor. De manera análoga, si queremos que el método de Runge-Kutta de orden cuatro sea mejor, deberá ofrecer una mayor precisión con el tamaño de paso h que el método de segundo orden con el tamaño de paso $\frac{1}{2}h$, porque el método de orden cuatro requiere el doble de evaluaciones por paso.

En el siguiente ejemplo se explica la superioridad del método de Runge-Kutta de cuarto orden según esta medida.

EJEMPLO 4 Para el problema

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

El método de Euler con $h = 0.025$, el método del punto medio con $h = 0.05$ y el método de Runge-Kutta de cuarto orden con $h = 0.1$ se comparan en los puntos de la red 0.1, 0.2, 0.3, 0.4 y 0.5. Todos requieren 20 evaluaciones funcionales para determinar los valores incluidos en la tabla 5.8 con que se aproxima $y(0.5)$. En este ejemplo, el método de cuarto orden resulta evidentemente superior. ■

Tabla 5.8

t_i	Valores exactos	Método de Euler $h = 0.025$	Método modificado de Euler $h = 0.05$	Método de Runge-Kutta de cuarto orden $h = 0.1$
0.0	0.5000000	0.5000000	0.5000000	0.5000000
0.1	0.6574145	0.6554982	0.6573085	0.6574144
0.2	0.8292986	0.8253385	0.8290778	0.8292983
0.3	1.0150706	1.0089334	1.0147254	1.0150701
0.4	1.2140877	1.2056345	1.2136079	1.2140869
0.5	1.4256394	1.4147264	1.4250141	1.4256384

CONJUNTO DE EJERCICIOS 5.4

1. Aplique el método modificado de Euler para aproximar las soluciones de los siguientes problemas de valor inicial y compare después los resultados con los valores reales.
   a. $y' = te^{3t} - 2y$, $0 \leq t \leq 1$, $y(0) = 0$, con $h = 0.5$; solución real $y(t) = \frac{1}{5}te^{3t} - \frac{1}{25}e^{3t} + \frac{1}{25}e^{-2t}$.
   b. $y' = 1+(t - y)^2$, $2 \leq t \leq 3$, $y(2) = 1$, con $h = 0.5$; solución real $y(t) = t + \frac{1}{1-t}$.
   c. $y' = 1 + y/t$, $1 \leq t \leq 2$, $y(1) = 2$, con $h = 0.25$; solución real $y(t) = t \ln t + 2t$.
   d. $y' = \cos 2t + \operatorname{sen} 3t$, $0 \leq t \leq 1$, $y(0) = 1$, con $h = 0.25$; solución real $y(t) = \frac{1}{2} \operatorname{sen} 2t - \frac{1}{3}\cos 3t + \frac{4}{3}$.
2. Repita el ejercicio 1 aplicando el método de Heun.
3. Repita el ejercicio 1 aplicando el método del punto medio.
4. Aplique el método modificado de Euler para aproximar las soluciones de los siguientes problemas de valor inicial y compare después los resultados con los valores reales.
   a. $y' = y/t - (y/t)^2$, $1 \leq t \leq 2$, $y(1) = 1$, con $h = 0.1$; solución real $y(t) = t/(1 + \ln t)$.
   b. $y' = 1 + y/t + (y/t)^2$, $1 \leq t \leq 3$, $y(1) = 0$, con $h = 0.2$; solución real $y(t) = t \tan (\ln t)$.
   c. $y' = -(y + 1)(y + 3)$, $0 \leq t \leq 2$, $y(0) = -2$, con $h = 0.2$; solución real $y(t) = -3 + 2(1 + e^{-2t})^{-1}$.
   d. $y' = -5y + 5t^2 + 2t$, $0 \leq t \leq 1$, $y(0) = \frac{1}{3}$, con $h = 0.1$; solución real $y(t) = t^2 + \frac{1}{3}e^{-5t}$.
5. Use los resultados del ejercicio 4 y la interpolación lineal para aproximar los valores de $y(t)$ y después compare los resultados con los valores reales.
   a. $y(1.25)$ y $y(1.93)$
   b. $y(2.1)$ y $y(2.75)$
   c. $y(1.3)$ y $y(1.93)$
   d. $y(0.54)$ y $y(0.94)$

6. Repita el ejercicio 4 aplicando el método de Heun.

7. Repita el ejercicio 5 usando los resultados del ejercicio 6.

8. Repita el ejercicio 4 aplicando el método del punto medio.

9. Repita el ejercicio 5 usando los resultados del ejercicio 8.

10. Repita el ejercicio 1 aplicando el método de Runge-Kutta de cuarto orden.

11. Repita el ejercicio 4 aplicando el método de Runge-Kutta de cuarto orden.

12. Use los resultados del ejercicio 11 y la interpolación cúbica de Hermite para aproximar los valores de $y(t)$ y compare las aproximaciones con los valores reales.

a. $y(1.25)$ y $y(1.93)$ **b.** $y(2.1)$ y $y(2.75)$

c. $y(1.3)$ y $y(1.93)$ **d.** $y(0.54)$ y $y(0.94)$

13. Demuestre que el método del punto medio, el método modificado de Euler y el método de Heun ofrecen las mismas aproximaciones al problema de valor real

$$y' = -y + t + 1, \quad 0 \le t \le 1, \quad y(0) = 1,$$

para cualquier elección de h. ¿Por qué es así?

14. Fluye agua de un tanque cónico invertido provisto de un orificio circular, con una velocidad

$$\frac{dx}{dt} = -0.6\pi r^2 \sqrt{2g}\,\frac{\sqrt{x}}{A(x)},$$

donde r es el radio del orificio, x es la altura del nivel del líquido medido desde el vértice del cono y $A(x)$ es el área de la sección transversal del tanque, a x unidades por arriba del orificio. Suponga que $r = 0.1$ pies, $g = 32.1$ pies/s^2, y que el tanque tiene un nivel inicial de agua de 8 pies y un volumen inicial de $512(\pi/3)$ pies^3.

a. Calcule el nivel del agua después de 10 min con $h = 20$ s.

b. Determine, con una exactitud de 1 min, cuándo se vaciará el tanque.

15. La reacción química irreversible en la cual dos moléculas de dicromato sólido de potasio ($K_2Cr_2O_7$), dos moléculas de agua (H_2O) y tres átomos de azufre sólido (S) se combinan para producir tres moléculas de dióxido gaseoso de azufre (SO_2), cuatro moléculas de hidróxido sólido de potasio (KOH) y dos moléculas de óxido sólido de cromo (Cr_2O_3), puede representarse simbólicamente por la *ecuación estequiométrica*:



$$2K_2Cr_2O_7 + 2H_2O + 3S \rightarrow 4KOH + 2Cr_2O_3 + 3SO_2.$$

Si originalmente se dispone de n_1 moléculas de $K_2Cr_2O_7$, n_2 moléculas de H_2O y n_3 moléculas de S, la siguiente ecuación diferencial describe la cantidad $x(t)$ de KOH después del tiempo t:

$$\frac{dx}{dt} = k\left(n_1 - \frac{x}{2}\right)^2\left(n_2 - \frac{x}{2}\right)^2\left(n_3 - \frac{3x}{4}\right)^3,$$

donde k es la constante de velocidad de la reacción. Si $k = 6.22 \times 10^{-19}$, $n_1 = n_2 = 2 \times 10^3$ y $n_3 = 3 \times 10^3$, ¿cuántas unidades de hidróxido de potasio se formarán después de 0.2 s?

16. Demuestre que el método de diferencias

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + a_1 f(t_i, w_i) + a_2 f(t_i + \alpha_2, w_i + \delta_2 f(t_i, w_i)),$$

para cada $i = 0, 1, \ldots, N-1$, no puede tener el error local de truncamiento $O(h^3)$ para cualquier elección de las constantes a_1, a_2, α_2 y δ_2.

17. El método de Runge-Kutta de cuarto orden puede escribirse en la forma

$$
\begin{aligned}
w_0 &= \alpha, \\
w_{i+1} &= w_i + \frac{h}{6} f(t_i, w_i) + \frac{h}{3} f(t_i + \alpha_1 h, w_i + \delta_1 h f(t_i, w_i)) \\
&\quad + \frac{h}{3} f(t_i + \alpha_2 h, w_i + \delta_2 h f(t_i + \gamma_2 h, w_i + \gamma_3 h f(t_i, w_i))) \\
&\quad + \frac{h}{6} f(t_i + \alpha_3 h, w_i + \delta_3 h f(t_i + \gamma_4 h, w_i + \gamma_5 h f(t_i + \gamma_6 h, w_i + \gamma_7 h f(t_i, w_i)))).
\end{aligned}
$$

Obtenga los valores de las constantes

$$\alpha_1, \alpha_2, \alpha_3, \delta_1, \delta_2, \delta_3, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6 \text{ y } \gamma_7.$$

5.5 Control del error y el método de Runge-Kutta-Fehlberg

En la sección 4.6 vimos el uso apropiado del tamaño variable de paso para producir métodos de aproximación de la integral, con la eficiencia requerida en la cantidad de cálculos. Por sí mismo, ello tal vez no sería suficiente para preferirlos, dada la mayor complejidad de su uso. Pero presentan otra característica que los hace sumamente útiles. En el procedimiento del tamaño de paso incorporan una estimación del error de truncamiento que no requiere aproximar las derivadas superiores de la función. A estos métodos se les llama *adaptativos*, porque adaptan el número y la posición de los nodos con que se efectúa la aproximación, para garantizar que el error de truncamiento no rebase la cota especificada.

Entre el problema de aproximar el valor de una integral definida y el aproximar la solución de un problema de valor inicial existe una estrecha relación. Así pues, no debe sorprendernos que haya métodos adaptativos que aproximan las soluciones de los problemas de valor inicial y que no sólo sean eficientes, sino que además incluyan el control de error.

Un método ideal de la ecuación de diferencias

$$w_{i+1} = w_i + h_i \phi(t_i, w_i, h_i), \qquad i = 0, 1, \ldots, N-1,$$

para aproximar la solución $y(t)$ al problema de valor inicial

$$y' = f(t, y), \qquad a \leq t \leq b, \qquad y(a) = \alpha,$$

deberá tener la propiedad de que, con una tolerancia $\varepsilon > 0$, la cantidad mínima de puntos de red servirá para asegurarse de que el error global $|y(t_i) - w_i|$, no rebasará ε con cualquier $i = 0, 1, \ldots, N$. No debe sorprendernos que tener una cantidad mínima de puntos de red y el control del error global de un método de diferencias, sea incompatible con el espaciamiento uniforme de los puntos en los intervalos. En esta sección estudiaremos los métodos con que se controla eficientemente el error de un método de ecuación de diferencias mediante la selección apropiada de los puntos de red.

Aunque no es posible, por lo general, determinar el error global de un método, en la sección 5.10 veremos que existe una estrecha relación entre el error local de truncamiento

y el error global. Mediante métodos de orden distinto podemos predecir el error local de truncamiento y seleccionar con esta predicción un tamaño de paso que controle el error global.

Para ilustrar este método, supongamos que tenemos dos métodos de aproximación. El primero es un método de n-ésimo orden obtenido de un método de Taylor de n-ésimo orden de la forma

$$y(t_{i+1}) = y(t_i) + h\phi(t_i, y(t_i), h) + O(h^{n+1}),$$

que produce las aproximaciones

$$\begin{aligned} w_0 &= \alpha \\ w_{i+1} &= w_i + h\phi(t_i, w_i, h), \qquad \text{para } i > 0, \end{aligned}$$

con el error de truncamiento $\tau_{i+1}(h) = O(h^n)$. En general, el método es generado al aplicar la modificación de Runge-Kutta al método de Taylor, pero la derivación específica carece de importancia.

El segundo método es similar, pero posee un orden mayor; se deriva del método de Taylor de orden $(n + 1)$-ésimo de la forma

$$y(t_{i+1}) = y(t_i) + h\tilde{\phi}(t_i, y(t_i), h) + O(h^{n+2}),$$

Se obtienen las siguientes aproximaciónes

$$\begin{aligned} \tilde{w}_0 &= a \\ \tilde{w}_{i+1} &= \tilde{w}_i + h\tilde{\phi}(t_i, \tilde{w}_i, h), \qquad \text{para } i > 0, \end{aligned}$$

con un error local de truncamiento $\tilde{\tau}_{i+1}(h) = O(h^{n+1})$.

Primero suponemos que $w_i \approx y(t_i) \approx \tilde{w}_i$ y seleccionamos un tamaño de paso fijo h para generar las aproximaciones w_{i+1} y $\tilde{w}_{i+1}$ a $y(t_{i+1})$. Entonces

$$\begin{aligned} \tau_{i+1}(h) &= \frac{y(t_{i+1}) - y(t_i)}{h} - \phi(t_i, y(t_i), h) \\ &= \frac{y(t_{i+1}) - w_i}{h} - \phi(t_i, w_i, h) \\ &= \frac{y(t_{i+1}) - [w_i + h\phi(t_i, w_i, h)]}{h} \\ &= \frac{1}{h}\,(y(t_{i+1}) - w_{i+1}). \end{aligned}$$

De manera similar

$$\tilde{\tau}_{i+1}(h) = \frac{1}{h}(y(t_{i+1}) - \tilde{w}_{i+1}).$$

En consecuencia

$$\begin{aligned} \tau_{i+1}(h) &= \frac{1}{h}(y(t_{i+1}) - w_{i+1}) \\ &= \frac{1}{h}[(y(t_{i+1}) - \tilde{w}_{i+1}) + (\tilde{w}_{i+1} - w_{i+1})] \\ &= \tilde{\tau}_{i+1}(h) + \frac{1}{h}(\tilde{w}_{i+1} - w_{i+1}). \end{aligned}$$

Pero $\tau_{i+1}(h)$ es $O(h^n)$ y $\tilde{\tau}_{i+1}(h)$ es $O(h^{n+1})$, por lo cual la parte significativa de $\tau_{i+1}(h)$ debe provenir de

$$\frac{1}{h}(\tilde{w}_{i+1} - w_{i+1}).$$

Esto nos da una aproximación calculada fácilmente del error local de trucamiento del método $O(h^n)$:

$$\tau_{i+1}(h) \approx \frac{1}{h}(\tilde{w}_{i+1} - w_{i+1}).$$

Sin embargo, el objetivo no es sólo estimar el error local del truncamiento, sino ajustar además el tamaño de paso para mantenerlo dentro de una cota especificada. Para hacerlo, ahora se supone que como $\tau_{i+1}(h)$ es $O(h^n)$, existe un número K independiente de h

$$\tau_{i+1}(h) \approx Kh^n.$$

Después podemos estimar el error local de truncamiento producido al aplicar el método de n-ésimo orden con un nuevo tamaño de paso qh, usando las aproximaciones originales w_{i+1} y $\tilde{w}_{i+1}$:

$$\tau_{i+1}(qh) \approx K(qh)^n = q^n(Kh^n) \approx q^n\tau_{i+1}(h) \approx \frac{q^n}{h}(\tilde{w}_{i+1} - w_{i+1}).$$

Para establecer la cota de $\tau_{i+1}(qh)$ por ε, escogemos q tal que

$$\frac{q^n}{h}\left|\tilde{w}_{i+1} - w_{i+1}\right| \approx \left|\tau_{i+1}(qh)\right| \leq \varepsilon;$$

es decir, tal que

$$q \leq \left(\frac{\varepsilon h}{|\tilde{w}_{i+1} - w_{i+1}|}\right)^{1/n}.$$

Un método muy usado que utiliza esta desigualdad para controlar el error es el **método de Runge-Kutta-Fehlberg**. (Véase [Fe].) Éste consiste en emplear el método de Runge-Kutta con el error local de truncamiento de quinto orden,

$$\tilde{w}_{i+1} = w_i + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6,$$

para estimar el error local en un método de Runge-Kutta de cuarto orden dado por

$$w_{i+1} = w_i + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4104}k_4 - \frac{1}{5}k_5,$$

donde

$$k_1 = hf(t_i, w_i),$$
$$k_2 = hf\left(t_i + \frac{h}{4}, w_i + \frac{1}{4}k_1\right),$$

$$k_3 = hf\left(t_i + \frac{3h}{8}, w_i + \frac{3}{32}k_1 + \frac{9}{32}k_2\right),$$
$$k_4 = hf\left(t_i + \frac{12h}{13}, w_i + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right),$$
$$k_5 = hf\left(t_i + h, w_i + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right),$$
$$k_6 = hf\left(t_i + \frac{h}{2}, w_i - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2656}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right).$$

Una ventaja de este método consiste en que sólo se requieren seis evaluaciones de f por paso. Los métodos arbitrarios de Runge-Kutta de cuarto y quinto orden usados de manera conjunta requieren (véase la tabla 5.7 en la sección 5.4) al menos cuatro evaluaciones de f con el método de cuarto orden y seis más con el de quinto orden, lo cual nos da un total de, por lo menos, diez evaluaciones de funciones.

En la teoría del control del error, un valor inicial de h en el i-ésimo paso se usó para obtener los primeros valores de w_{i+1} y $\tilde{w}_{i+1}$, que nos permitieron determinar q en ese paso, y luego se repitieron los cálculos. Este procedimiento requiere el doble de evaluaciones de funciones por paso, sin control de error. En la práctica, el valor de q a usar se selecciona de manera un poco diferente, a fin de que valga la pena el aumento de evaluaciones de funciones. El valor de q determinado en el i-ésimo paso cumple dos propósitos:

1. Para rechazar, de ser necesario, la elección inicial de h en el paso i-ésimo y repetir los cálculos por medio de qh, y
2. Para predecir una elección adecuada de h para el $(i + 1)$-ésimo paso.

Debido a la cuota que debe pagarse en términos de evaluaciones de funciones si se repiten los pasos, q tiende a ser elegida de manera conservadora; de hecho, en el método de Runge-Kutta-Fehlberg con $n = 4$, la elección común es

$$q = \left(\frac{\varepsilon h}{2|\tilde{w}_{i+1} - w_{i+1}|}\right)^{1/4} = 0.84\left(\frac{\varepsilon h}{|\tilde{w}_{i+1} - w_{i+1}|}\right)^{1/4}.$$

En el algoritmo 5.3 para el método de Runge-Kutta-Fehlberg, se agrega el paso 9 para suprimir grandes modificaciones al tamaño del paso. Esto se hace para no tener que dedicar mucho tiempo a los tamaños pequeños de paso en las regiones donde hay irregularidades de las derivadas de y, y para evitar los grandes tamaños de paso, que pueden llevar a omitir las regiones sensibles entre los pasos. En algunos casos, en el algoritmo se omite totalmente el procedimiento que aumenta el tamaño del paso, y el procedimiento con que se disminuye el tamaño se modifica para que se incorpore sólo cuando es necesario controlar el error.

ALGORITMO 5.3

Método de Runge-Kutta-Fehlberg

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

con un error local de truncamiento que no rebase la tolerancia especificada:

ENTRADA extremos a, b; condición inicial α; tolerancia TOL; tamaño máximo de paso $hmáx$; tamaño mínimo de paso $hmín$.

SALIDA t, w, h donde w aproxima a $y(t)$ y se usó el tamaño de paso h o un mensaje de que se rebasó el tamaño mínimo de paso.

Paso 1 Tome $t = a$;
$w = \alpha$;
$h = hmáx$;
$BAND = 1$;
SALIDA (t, w).

Paso 2 Mientras ($BAND = 1$) haga pasos 3-11.

Paso 3 Tome $K_1 = hf(t, w)$;

$$K_2 = hf(t + \tfrac{1}{4}h, w + \tfrac{1}{4}K_1);$$
$$K_3 = hf(t + \tfrac{3}{8}h, w + \tfrac{3}{32}K_1 + \tfrac{9}{32}K_2);$$
$$K_4 = hf(t + \tfrac{12}{13}h, w + \tfrac{1932}{2197}K_1 - \tfrac{7200}{2197}K_2 + \tfrac{7296}{2197}K_3);$$
$$K_5 = hf(t + h, w + \tfrac{439}{216}K_1 - 8K_2 + \tfrac{3680}{513}K_3 - \tfrac{845}{4104}K_4);$$
$$K_6 = hf(t + \tfrac{1}{2}h, w - \tfrac{8}{27}K_1 + 2K_2 - \tfrac{3544}{2565}K_3 + \tfrac{1859}{4104}K_4 - \tfrac{11}{40}K_5).$$

Paso 4 Tome $R = \frac{1}{h}\left|\frac{1}{360}K_1 - \frac{128}{4275}K_3 - \frac{2197}{75240}K_4 + \frac{1}{50}K_5 + \frac{2}{55}K_6\right|$.

(*Nota*: $R = \frac{1}{h}|\tilde{w}_{i+1} - w_{i+1}|$.)

Paso 5 Si $R \leq TOL$ entonces haga pasos 6 y 7.

Paso 6 Tome $t = t + h$; (*aproximación aceptada*).

$$w = w + \tfrac{25}{216}K_1 + \tfrac{1408}{2565}K_3 + \tfrac{2197}{4104}K_4 - \tfrac{1}{5}K_5.$$

Paso 7 SALIDA (t, w, h).

Paso 8 Tome $\delta = 0.84(TOL/R)^{1/4}$.

Paso 9 Si $\delta \leq 0.1$ entonces tome $h = 0.1h$
o si $\delta \geq 4$ entonces tome $h = 4h$
de otro modo tome $h = \delta h$. (*Calcule de nuevo h.*)

Paso 10 Si $h > hmáx$ entonces tome $h = hmáx$.

Paso 11 Si $t \geq b$ entonces tome $BAND = 0$
de otro modo si $t + h > b$ entonces tome $h = b - t$
de otro modo si $h < hmín$ entonces
tome $BAND = 0$;
SALIDA ('*rebasado h mínimo*').
(*Procedimiento terminado de manera no satisfactoria.*)

Paso 12 (*El procedimiento se completó.*)
PARAR. ■

EJEMPLO 1 Utilizaremos el algoritmo 5.3 para aproximar la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

que tiene la solución $y(t) = (t + 1)^2 - 0.5e^t$. La entrada del algoritmo es la tolerancia $TOL = 10^{-5}$, un tamaño máximo de paso $hmáx = 0.25$ y un tamaño mínimo de paso $hmín = 0.01$. Los resultados se muestran en la tabla 5.9. Las dos últimas columnas de la tabla contienen los resultados del método de quinto orden. Con valores pequeños de t, el error es menor que el del método de cuarto orden, pero es mayor que cuando t aumenta. ∎

Tabla 5.9

		RKF-4				RKF-5	
t_i	$y_i = y(t_i)$	w_i	h_i	R_i	$\|y_i - w_i\|$	$\hat{w}_i$	$\|y_i - \hat{w}_i\|$
0	0.5	0.5			0.5		
0.2500000	0.9204873	0.9204886	0.2500000	6.2×10^{-6}	1.3×10^{-6}	0.9204870	2.424×10^{-7}
0.4865522	1.3964884	1.3964910	0.2365522	4.5×10^{-6}	2.6×10^{-6}	1.3964900	1.510×10^{-6}
0.7293332	1.9537446	1.9537488	0.2427810	4.3×10^{-6}	4.2×10^{-6}	1.9537477	3.136×10^{-6}
0.9793332	2.5864198	2.5864260	0.2500000	3.8×10^{-6}	6.2×10^{-6}	2.5864251	5.242×10^{-6}
1.2293332	3.2604520	3.2604605	0.2500000	2.4×10^{-6}	8.5×10^{-6}	3.2604599	7.895×10^{-6}
1.4793332	3.9520844	3.9520955	0.2500000	7×10^{-7}	1.11×10^{-5}	3.9520954	1.096×10^{-5}
1.7293332	4.6308127	4.6308268	0.2500000	1.5×10^{-6}	1.41×10^{-5}	4.6308272	1.446×10^{-5}
1.9793332	5.2574687	5.2574861	0.2500000	4.3×10^{-6}	1.73×10^{-5}	5.2574871	1.839×10^{-5}
2.0000000	5.3054720	5.3054896	0.0206668		1.77×10^{-5}	5.3054896	1.768×10^{-5}

Para ejecutar el método de Runge-Kutta-Fehlberg usando Maple, se utiliza el comando `dsolve` con la opción numérica. Considere el problema de valor inicial del ejemplo 1. El comando

```
>g:=dsolve({D(y)(t)=y(t)-t*t+1,y(0)=0.5},y(t),numeric);
```

devuelve el procedimiento

$$g := \text{proc}(rkf45_x) \ldots \text{ end}$$

Como se indica en el ejemplo, podemos evaluar y por medio de

```
>g(2.0);
```

que nos da

$$[t = 2.0, y(t) = 5.305471958400194]$$

CONJUNTO DE EJERCICIOS 5.5

1. Aplique el método de Runge-Kutta-Fehlberg con la tolerancia $TOL = 10^{-4}$, $hmáx = 0.25$ y $hmín = 0.05$ para aproximar las soluciones de los siguientes problemas de valor inicial. Después, compare los resultados con los valores reales.

a. $y' = y/t - (y/t)^2$, $0 \leq t \leq 1$, $y(0) = 0$; solución real $y(t) = \frac{1}{5}te^{3t} - \frac{1}{25}e^{3t} + \frac{1}{25}e^{-2t}$.
b. $y' = 1 + (t - y)^2$, $2 \leq t \leq 3$, $y(2) = 1$; solución real $y(t) = t + 1/(1 - t)$.
c. $y' = 1 + y/t$, $1 \leq t \leq 2$, $y(1) = 2$; solución real $y(t) = t \ln t + 2t$.
d. $y' = \cos 2t + \operatorname{sen} 3t$, $0 \leq t \leq 1$, $y(0) = 1$; solución real $y(t) = \frac{1}{2} \operatorname{sen} 2t - \frac{1}{3} \cos 3t + \frac{4}{3}$.

2. Use el algoritmo de Runge-Kutta-Fehlberg con la tolerancia $TOL = 10^{-4}$ para aproximar la solución de los siguientes problemas de valor inicial.
a. $y' = (y/t)^2 + y/t$, $1 \leq t \leq 1.2$, $y(1) = 1$, con $hmáx = 0.005$ y $hmín = 0.02$.
b. $y' = \operatorname{sen} t + e^{-t}$, $0 \leq t \leq 1$, $y(0) = 0$, con $hmáx = 0.25$ y $hmín = 0.02$.
c. $y' = 1/t(y^2 + y)$, $1 \leq t \leq 3$, $y(1) = -2$, con $hmáx = 0.5$ y $hmín = 0.02$.
d. $y' = t^2$, $0 \leq t \leq 2$, $y(0) = 0$, con $hmáx = 0.5$ y $hmín = 0.02$.

3. Aplique el método de Runge-Kutta-Fehlberg con la tolerancia $TOL = 10^{-6}$, $hmáx = 0.5$ y $hmín = 0.05$ para aproximar las soluciones de los siguientes problemas de valor inicial. Después, compare los resultados con los valores reales.
a. $y' = y/t - (y/t)^2$, $1 \leq t \leq 4$, $y(1) = 1$; solución real $y(t) = t/(1 + \ln t)$.
b. $y' = 1 + y/t + (y/t)^2$, $1 \leq t \leq 3$, $y(1) = 0$; solución real $y(t) = t \tan(\ln t)$.
c. $y' = -(y + 1)(y + 3)$, $0 \leq t \leq 3$, $y(0) = -2$; solución real $y(t) = -3 + 2(1 + e^{-2t})^{-1}$.
d. $y' = (t + 2t^3)y^3 - ty$, $0 \leq t \leq 2$, $y(0) = \frac{1}{3}$; solución real $y(t) = (3 + 2t^2 + 6e^{t^2})^{-1/2}$.

4. El método de Runge-Kutta-Verner se basa en las fórmulas

$$w_{i+1} = w_i + \frac{13}{160}k_1 + \frac{2375}{5984}k_3 + \frac{5}{16}k_4 + \frac{12}{85}k_5 + \frac{3}{44}k_6 \text{ y}$$

$$\tilde{w}_{i+1} = w_i + \frac{3}{40}k_1 + \frac{875}{2244}k_3 + \frac{23}{72}k_4 + \frac{264}{1955}k_5 + \frac{125}{11592}k_7 + \frac{43}{616}k_8,$$

donde

$$k_1 = hf(t_i, w_i),$$

$$k_2 = hf\left(t_i + \frac{h}{6}, w_i + \frac{1}{6}k_1\right),$$

$$k_3 = hf\left(t_i + \frac{4h}{15}, w_i + \frac{4}{75}k_1 + \frac{16}{75}k_2\right),$$

$$k_4 = hf\left(t_i + \frac{2h}{3}, w_i + \frac{5}{6}k_1 - \frac{8}{3}k_2 + \frac{5}{2}k_3\right),$$

$$k_5 = hf\left(t_i + \frac{5h}{6}, w_i - \frac{165}{64}k_1 + \frac{55}{6}k_2 - \frac{425}{64}k_3 + \frac{85}{96}k_4\right),$$

$$k_6 = hf\left(t_i + h, w_i + \frac{12}{5}k_1 - 8k_2 + \frac{4015}{612}k_3 - \frac{11}{36}k_4 + \frac{88}{255}k_5\right),$$

$$k_7 = hf\left(t_i + \frac{h}{15}, w_i - \frac{8263}{15000}k_1 + \frac{124}{75}k_2 - \frac{643}{680}k_3 - \frac{81}{250}k_4 + \frac{2484}{10625}k_5\right),$$

$$k_8 = hf\left(t_i + h, w_i + \frac{3501}{1720}k_1 - \frac{300}{43}k_2 + \frac{297275}{52632}k_3 - \frac{319}{2322}k_4 + \frac{24068}{84065}k_5 + \frac{3850}{26703}k_7\right).$$

El método de sexto orden $\tilde{w}_{i+1}$ sirve para estimar el error en el método de quinto orden w_{i+1}. Construya un algoritmo semejante al de Runge-Kutta-Fehlberg y repita el ejercicio 3 usando este nuevo método.

5. En la teoría de la propagación de enfermedades contagiosas (véase [Ba1] o [Ba2]), podemos utilizar una ecuación diferencial relativamente elemental para predecir el número de individuos de la población infectados en un tiempo dado, siempre y cuando realicemos las suposiciones de simplificación adecuadas. En particular, supongamos que todos los individuos de una población fija tienen la misma probabilidad de infectarse y que, una vez infectados, permanecen en ese estado. Si con $x(t)$ denotamos al número de individuos vulnerables en el tiempo t y si con $y(t)$ denotamos al número de los infectados, podemos suponer, razonablemente, que la rapidez con que el número de los infectados cambia es proporcional al producto de $x(t)$ y $y(t)$, porque la rapidez depende del número de individuos infectados y del número de individuos vulnerables que existen en ese tiempo. Si la población es lo suficientemente numerosa para suponer que $x(t)$ y $y(t)$ son variables continuas, podemos expresar el problema como

$$y'(t) = kx(t)y(t),$$

donde k es una constante y $x(t) + y(t) = m$ es la población total. Podemos reescribir esta ecuación para que contenga sólo $y(t)$ como

$$y'(t) = k(m - y(t))y(t).$$

a. Suponiendo que $m = 100\,000$, $y(0) = 1000$, $k = 2 \times 10^{-6}$, y que el tiempo se mide en días, encuentre una aproximación al número de individuos infectados al cabo de 30 días.

b. La ecuación diferencial del inciso (a) se denomina *ecuación de Bernoulli* y puede transformarse en una ecuación diferencial lineal en $u(t)$ tomando $u(t) = (y(t))^{-1}$. Aplique este método para encontrar la solución exacta de la ecuación, bajo los mismos supuestos del inciso (a); después, compare el valor vedadero de $y(t)$ en la aproximación aquí dada. ¿Qué es $\lim_{t\to\infty} y(t)$? ¿Concuerda esto con lo que usted intuye?

6. En el ejercicio anterior, todos los individuos infectados permanecieron en la población y propagaron la enfermedad. Una propuesta más realista consiste en introducir una tercera variable $z(t)$, que representa el número de las personas a quienes en un tiempo dado t se les separa de la población infectada por aislamiento, recuperación y la subsecuente inmunidad o fallecimiento. Naturalmente esto viene a complicar el problema, pero podemos demostrar (véase [Ba2]) que se puede obtener una solución aproximada en la forma

$$x(t) = x(0)e^{-(k_1/k_2)z(t)} \quad \text{y} \quad y(t) = m - x(t) - z(t),$$

donde k_1 es la rapidez de la infección, k_2 es la rapidez del aislamiento y $z(t)$ se obtiene de la ecuación diferencial

$$z'(t) = k_2(m - z(t) - x(0)e^{-(k_1/k_2)z(t)}).$$

Los autores no conocen método alguno para resolver directamente este problema y, por lo mismo, es necesario aplicar un procedimiento numérico. Obtenga una aproximación a $z(30)$, a $y(30)$ y a $x(30)$, suponiendo que $m = 100\,000$, $x(0) = 99\,000$, $k_1 = 2 \times 10^{-6}$, y que $k_2 = 10^{-4}$.

5.6 Métodos multipasos

Los métodos que hemos explicado hasta ahora se llaman **métodos de un paso**, porque la aproximación del punto de red t_{i+1} contiene información proveniente de uno solo de los puntos anteriores de red t_i. Aunque estas técnicas pueden usar la información relativa a la

evaluación de funciones en los puntos entre t_i y t_{i+1}, no la conservan para utilizarla directamente en aproximaciones futuras. Toda la información que emplean se obtiene dentro del subintervalo en que va a aproximarse la solución.

Como la solución aproximada está disponible en los puntos de red $t_0, t_1, \ldots, t_i$ antes de obtener la aproximación en t_{i+1} y como el error $|w_j - y(t_j)|$ tiende a aumentar con j, parece razonable desarrollar métodos que usen estos datos precedentes más precisos al aproximar la solución en t_{i+1}.

Se conoce como métodos *multipasos* a aquellos que emplean la aproximación en más de uno de los puntos de red precedentes para determinar la aproximación en el siguiente punto. A continuación se da la definición exacta de estos métodos, junto con la de dos tipos de ellos.

Definición 5.14 Un **método multipasos de paso *m*** para resolver el problema de valor inicial

$$y' = f(t, y), \qquad a \le t \le b, \qquad y(a) = \alpha, \tag{5.22}$$

es aquel cuya ecuación de diferencias para obtener la aproximación w_{i+1} en el punto de red t_{i+1} puede representarse por medio de la siguiente ecuación, donde m es un entero mayor que 1:

$$\begin{aligned} w_{i+1} = {} & a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} \\ & + h[b_m f(t_{i+1}, w_{i+1}) + b_{m-1} f(t_i, w_i) \\ & + \cdots + b_0 f(t_{i+1-m}, w_{i+1-m})], \end{aligned} \tag{5.23}$$

para $i = m-1, m, \ldots, N-1$, donde $h = (b-a)/N$, $a_0, a_1, \ldots, a_{m-1}$ y $b_0, b_1, \ldots, b_m$ son constantes y los valores iniciales

$$w_0 = \alpha, \qquad w_1 = \alpha_1, \qquad w_2 = \alpha_2, \; \ldots, \; w_{m-1} = \alpha_{m-1}$$

son especificados. ■

Cuando $b_m = 0$, el método es **explícito o abierto**, ya que la ecuación (5.23) da entonces w_{i+1} de manera explícita en términos de los valores previamente determinados. Cuando $b_m \neq 0$, el método es **implícito o cerrado**, ya que w_{i+1} se encuentra en ambos lados de la ecuación (5.23) y se especifica sólo implícitamente.

EJEMPLO 1 Las ecuaciones

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3, \tag{5.24}$$

$$w_{i+1} = w_i + \frac{h}{24}[55 f(t_i, w_i) - 59 f(t_{i-1}, w_{i-1}) + 37 f(t_{i-2}, w_{i-2}) - 9 f(t_{i-3}, w_{i-3})],$$

para cada $i = 3, 4, \ldots, N-1$, definen un método explícito de cuatro pasos llamado **método de Adams-Bashforth de cuarto orden.** Las ecuaciones

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \tag{5.25}$$

$$w_{i+1} = w_i + \frac{h}{24}[9 f(t_{i+1}, w_{i+1}) + 19 f(t_i, w_i) - 5 f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})],$$

para cada $i = 2, 3, \ldots, N - 1$, definen un método implícito de tres pasos denominado **método de Adams-Moulton de cuarto orden.** ■

En (5.24) o en (5.25) deben especificarse los valores iniciales, generalmente suponiendo que $w_0 = \alpha$ y generando los valores restantes por medio de un método de Runge-Kutta o bien con otro método de un paso.

Si queremos aplicar directamente un método implícito como el (5.25), debemos resolver la ecuación implícita para w_{i+1}. No es evidente que podamos hacer esto en general, ni que siempre obtendremos una solución única para w_{i+1}.

Antes de comenzar la deducción de los métodos multipasos, observe que la solución del problema de valor inicial (5.22), si lo integramos en el intervalo $[t_i, t_{i+1}]$, tiene la propiedad de que

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(t)\, dt = \int_{t_i}^{t_{i+1}} f(t, y(t))\, dt.$$

En consecuencia,

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(t, y(t))\, dt. \tag{5.26}$$

Como no podemos integrar $f(t, y(t))$ sin conocer $y(t)$, que es la solución del problema, en lugar de ello integramos un polinomio interpolante $P(t)$ a $f(t, y(t))$ que se determina con algunos de los puntos de datos obtenidos previamente $(t_0, w_0), (t_1, w_1), \ldots, (t_i, w_i)$. Cuando, además, suponemos que $y(t_i) \approx w_i$, la ecuación (5.26) se convierte en

$$y(t_{i+1}) \approx w_i + \int_{t_i}^{t_{i+1}} P(t)\, dt. \tag{5.27}$$

Aunque en la deducción podemos utilizar cualquier forma del polinomio interpolante, lo más adecuado es emplear la fórmula de diferencia regresiva de Newton.

Para derivar un método explícito de Adams-Bashforth de m pasos, formamos el polinomio de diferencias regresivas $P_{m-1}(t)$ a través de $(t_i, f(t_i, y(t_i))), (t_{i-1}, f(t_{i-1}, y(t_{i-1}))), \ldots, (t_{i+1-m}, f(t_{i+1-m}, y(t_{i+1-m})))$. Puesto que $P_{m-1}(t)$ es un polinomio interpolante de grado $m - 1$, existe un número ξ_i en (t_{i+1-m}, t_i) con

$$f(t, y(t)) = P_{m-1}(t) + \frac{f^{(m)}(\xi_i, y(\xi_i))}{m!}(t - t_i)(t - t_{i-1}) \cdots (t - t_{i+1-m}).$$

La introducción de la sustitución de la variable $t = t_i + sh$ con $dt = h\, ds$ en $P_{m-1}(t)$ y el término de error implica que

$$\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt = \int_{t_i}^{t_{i+1}} \sum_{k=0}^{m-1} (-1)^k \binom{-s}{k} \nabla^k f(t_i, y(t_i))\, dt$$
$$+ \int_{t_i}^{t_{i+1}} \frac{f^{(m)}(\xi_i, y(\xi_i))}{m!}(t - t_i)(t - t_{i-1}) \cdots (t - t_{i+1-m})\, dt$$

$$= \sum_{k=0}^{m-1} \nabla^k f(t_i, y(t_i))h(-1)^k \int_0^1 \binom{-s}{k} ds$$

$$+ \frac{h^{m+1}}{m!} \int_0^1 s(s+1)\cdots(s+m-1) f^{(m)}(\xi_i, y(\xi_i))\, ds.$$

Las integrales $(-1)^k \int_0^1 \binom{-s}{k} ds$ para diversos valores de k son fáciles de evaluar y se incluyen en la tabla 5.10. Por ejemplo, cuando $k = 3$,

$$(-1)^3 \int_0^1 \binom{-s}{3} ds = -\int_0^1 \frac{(-s)(-s-1)(-s-2)}{1 \cdot 2 \cdot 3} ds$$

$$= \frac{1}{6}\int_0^1 (s^3 + 3s^2 + 2s)\, ds$$

$$= \frac{1}{6}\left[\frac{s^4}{4} + s^3 + s^2\right]_0^1 = \frac{1}{6}\left(\frac{9}{4}\right) = \frac{3}{8}.$$

Tabla 5.10

k	0	1	2	3	4	5
$(-1)^k \int_0^1 \binom{-s}{k} ds$	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$

En consecuencia,

$$\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt = h\left[f(t_i, y(t_i)) + \frac{1}{2}\nabla f(t_i, y(t_i)) + \frac{5}{12}\nabla^2 f(t_i, y(t_i)) + \cdots\right]$$

$$+ \frac{h^{m+1}}{m!}\int_0^1 s(s+1)\cdots(s+m-1) f^{(m)}(\xi_i, y(\xi_i))\, ds. \tag{5.28}$$

Puesto que $s(s+1)\cdots(s+m-1)$ no cambia de signo en $[0, 1]$, podemos aplicar el teorema del valor medio ponderado de las integrales y deducir que para algún número μ_i, donde $t_{i+1-m} < \mu_i < t_{i+1}$, el término de error de la ecuación (5.28) se convierte en

$$\frac{h^{m+1}}{m!}\int_0^1 s(s+1)\cdots(s+m-1) f^{(m)}(\xi_i, y(\xi_i))\, ds$$

$$= \frac{h^{m+1} f^{(m)}(\mu_i, y(\mu_i))}{m!}\int_0^1 s(s+1)\cdots(s+m-1)\, ds$$

o en

$$h^{m+1} f^{(m)}(\mu_i, y(\mu_i))(-1)^m \int_0^1 \binom{-s}{m} ds. \tag{5.29}$$

Dado que $y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(t, y(t))\, dt$, podemos escribir la ecuación (5.28) así:

$$y(t_{i+1}) = y(t_i) + h\left[f(t_i, y(t_i)) + \frac{1}{2}\nabla f(t_i, y(t_i)) + \frac{5}{12}\nabla^2 f(t_i, y(t_i)) + \cdots\right]$$

$$+ h^{m+1} f^{(m)}(\mu_i, y(\mu_i))(-1)^m \int_0^1 \binom{-s}{m} ds. \tag{5.30}$$

EJEMPLO 2 Para derivar el método de Adams-Bashforth de tres pasos, consideremos la ecuación (5.30) con $m = 3$:

$$\begin{aligned}
y(t_{i+1}) &= y(t_i) + h\left[f(t_i, y(t_i)) + \frac{1}{2}\nabla f(t_i, y(t_i)) + \frac{5}{12}\nabla^2 f(t_i, y(t_i))\right] \\
&= y(t_i) + h\left\{f(t_i, y(t_i)) + \frac{1}{2}[f(t_i, y(t_i)) - f(t_{i-1}, y(t_{i-1}))]\right. \\
&\quad \left. + \frac{5}{12}[f(t_i, y(t_i)) - 2f(t_{i-1}, y(t_{i-1})) + f(t_{i-2}, y(t_{i-2}))]\right\} \\
&= y(t_i) + \frac{h}{12}[23 f(t_i, y(t_i)) - 16 f(t_{i-1}, y(t_{i-1})) + 5 f(t_{i-2}, y(t_{i-2}))].
\end{aligned}$$

Por tanto, el método de Adams-Bashforth de tres pasos es

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2,$$
$$w_{i+1} = w_i + \frac{h}{12}[23 f(t_i, w_i) - 16 f(t_{i-1}, w_{i-1})] + 5 f(t_{i-2}, w_{i-2})],$$

para $i = 2, 3, \ldots, N-1$. ■

También podemos derivar los métodos multipasos por medio de la serie de Taylor. En el ejercicio 10 se incluye un ejemplo del procedimiento en cuestión. En el ejercicio 9 se explica una derivación utilizando el polinomio interpolante de Lagrange.

En los métodos multipasos el error local de truncamiento se define de manera análoga al de los métodos de un paso. Al igual que en el caso de estos últimos, el error local de truncamiento ofrece una medida de cómo la solución de la ecuación diferencial no logra resolver la ecuación de diferencias.

Definición 5.15 Si $y(t)$ es la solución al problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

y si

$$\begin{aligned}
w_{i+1} &= a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} \\
&\quad + h[b_m f(t_{i+1}, w_{i+1}) + b_{m-1} f(t_i, w_i) + \cdots + b_0 f(t_{i+1-m}, w_{i+1-m})]
\end{aligned}$$

es el $(i+1)$-ésimo paso en un método multipasos, el **error local de truncamiento** en este paso será

$$\tau_{i+1}(h) = \frac{y(t_{i+1}) - a_{m-1}y(t_i) - \cdots - a_0 y(t_{i+1-m})}{h} \tag{5.31}$$

$$- [b_m f(t_{i+1}, y(t_{i+1})) + \cdots + b_0 f(t_{i+1-m}, y(t_{i+1-m}))],$$

para cada $i = m-1, m, \ldots, N-1$. ■

EJEMPLO 3 Para determinar el error local de truncamiento en el método de Adams-Bashforth de tres pasos obtenido en el ejemplo 2, consideremos la forma del error dada en la ecuación (5.29) y el dato correspondiente en la tabla 5.10:

$$h^4 f^{(3)}(\mu_i, y(\mu_i))(-1)^3 \int_0^1 \binom{-s}{3} ds = \frac{3h^4}{8} f^{(3)}(\mu_i, y(\mu_i)).$$

Al aplicar el hecho de que $f^{(3)}(\mu_i, y(\mu_i)) = y^{(4)}(\mu_i)$ y la ecuación de diferencias deducida en el ejemplo 2, tenemos

$$\tau_{i+1}(h) = \frac{y(t_{i+1}) - y(t_i)}{h} - \frac{1}{12}[23 f(t_i, y(t_i)) - 16 f(t_{i-1}, y(t_{i-1})) + 5 f(t_{i-2}, y(t_{i-2}))]$$

$$= \frac{1}{h}\left[\frac{3h^4}{8} f^{(3)}(\mu_i, y(\mu_i))\right] = \frac{3h^3}{8} y^{(4)}(\mu_i), \qquad \text{para alguna } \mu_i \in (t_{i-2}, t_{i+1}). \quad ■$$

A continuación se incluyen algunos de los métodos multipasos explícitos, junto con sus valores iniciales requeridos y los errores locales de truncamiento. La deducción de estos métodos es semejante al procedimiento de los ejemplos 2 y 3.

Método explícito de Adams-Bashforth de dos pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1,$$

$$w_{i+1} = w_i + \frac{h}{2}[3 f(t_i, w_i) - f(t_{i-1}, w_{i-1})], \tag{5.32}$$

donde $i = 1, 2, \ldots, N-1$. El error local de truncamiento $\tau_{i+1}(h) = \frac{5}{12} y'''(\mu_i)h^2$, para alguna $\mu_i \in (t_{i-1}, t_{i+1})$.

Método explícito de Adams-Bashforth de tres pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2,$$

$$w_{i+1} = w_i + \frac{h}{12}[23 f(t_i, w_i) - 16 f(t_{i-1}, w_{i-1}) + 5 f(t_{i-2}, w_{i-2})], \tag{5.33}$$

donde $i = 2, 3, \ldots, N-1$. El error local de truncamiento es $\tau_{i+1}(h) = \frac{3}{8} y^{(4)}(\mu_i)h^3$, para algún $\mu_i \in (t_{i-2}, t_{i+1})$.

Método explícito de Adams-Bashforth de cuatro pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3, \tag{5.34}$$

$$w_{i+1} = w_i + \frac{h}{24}[55 f(t_i, w_i) - 59 f(t_{i-1}, w_{i-1}) + 37 f(t_{i-2}, w_{i-2}) - 9 f(t_{i-3}, w_{i-3})],$$

donde $i = 3, 4, \ldots, N-1$. El error local de truncamiento es $\tau_{i+1}(h) = \frac{251}{720} y^{(5)}(\mu_i)h^4$, para algún $\mu_i \in (t_{i-3}, t_{i+1})$.

Método explícito de Adams-Bashforth de cinco pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3, \quad w_4 = \alpha_4,$$

$$w_{i+1} = w_i + \frac{h}{720}[1901 f(t_i, w_i) - 2774 f(t_{i-1}, w_{i-1}) \tag{5.35}$$

$$+ 2616 f(t_{i-2}, w_{i-2}) - 1274 f(t_{i-3}, w_{i-3}) + 251 f(t_{i-4}, w_{i-4})],$$

donde $i = 4, 5, \ldots, N - 1$. El error local de truncamiento es $\tau_{i+1}(h) = \frac{95}{288} y^{(6)}(\mu_i)h^5$, para algún $\mu_i \in (t_{i-4}, t_{i+1})$.

Los métodos implícitos se obtienen utilizando $(t_{i+1}, f(t_{i+1}, y(t_{i+1})))$ como nodo de interpolación adicional en la aproximación de la integral

$$\int_{t_i}^{t_{i+1}} f(t, y(t))\, dt.$$

A continuación se incluyen algunos de los métodos implícitos más comunes.

Método implícito de Adams-Moulton de tres pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1,$$

$$w_{i+1} = w_i + \frac{h}{12}[5 f(t_{i+1}, w_{i+1}) + 8 f(t_i, w_i) - f(t_{i-1}, w_{i-1})], \tag{5.36}$$

donde $i = 1, 2, \ldots, N - 1$. El error local de truncamiento es $\tau_{i+1}(h) = -\frac{1}{24} y^{(4)}(\mu_i)h^3$, para algún $\mu_i \in (t_{i-1}, t_{i+1})$.

Método implícito de Adams Moulton de tres pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \tag{5.37}$$

$$w_{i+1} = w_i + \frac{h}{24}[9 f(t_{i+1}, w_{i+1}) + 19 f(t_i, w_i) - 5 f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})],$$

donde $i = 1, 2, \ldots, N - 1$. El error local de truncamiento es $\tau_{i+1}(h) = -\frac{19}{720} y^{(5)}(\mu_i)h^4$, para algún $\mu_i \in (t_{i-2}, t_{i+1})$.

Método implícito de Adams-Moulton de cuatro pasos:

$$w_0 = \alpha, \quad w_1 = \alpha_1, \quad w_2 = \alpha_2, \quad w_3 = \alpha_3,$$

$$w_{i+1} = w_i + \frac{h}{720}[251 f(t_{i+1}, w_{i+1}) + 646 f(t_i, w_i) \tag{5.38}$$

$$-264 f(t_{i-1}, w_{i-1}) + 106 f(t_{i-2}, w_{i-2}) - 19 f(t_{i-3}, w_{i-3})],$$

donde $i = 3, 4, \ldots, N - 1$. El error local de truncamiento es $\tau_{i+1}(h) = -\frac{3}{160} y^{(6)}(\mu_i)h^5$ para algún $\mu_i \in (t_{i-3}, t_{i+1})$.

Es interesante comparar un método explícito de Adams-Bashforth de m pasos con un método implícito de Adams-Moulton de $(m - 1)$ pasos. Ambos requieren m evaluaciones

de f por paso y tienen los términos $y^{(m+1)}(\mu_i)h^m$ en sus errores locales de truncamiento. En términos generales, los coeficientes de los términos que contienen f en el error local de truncamiento son menores en los métodos implícitos que en los explícitos. Esto da origen a una mayor estabilidad y a menores errores de redondeo en los métodos implícitos.

EJEMPLO 4 Consideremos el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

y las aproximaciones dadas por el método explícito de Adams-Bashforth de cuatro pasos y el implícito de Adams-Moulton de tres pasos, en ambos con $h = 0.2$.

El método de Adams-Bashforth tiene la ecuación de diferencias

$$w_{i+1} = w_i + \frac{h}{24}[55f(t_i, w_i) - 59f(t_{i-1}, w_{i-1}) + 37f(t_{i-2}, w_{i-2}) - 9f(t_{i-3}, w_{i-3})],$$

para $i = 3, 4, \ldots, 9$, que cuando se simplifica mediante $f(t, y) = y - t^2 + 1$, $h = 0.2$ y mediante $t_i = 0.2i$, se convierte en

$$w_{i+1} = \frac{1}{24}[35w_i - 11.8w_{i-1} + 7.4w_{i-2} - 1.8w_{i-3} - 0.192i^2 - 0.192i + 4.736].$$

El método de Adams-Moulton tiene la ecuación de diferencias

$$w_{i+1} = w_i + \frac{h}{24}[9f(t_{i+1}, w_{i+1}) + 19f(t_i, w_i) - 5f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})],$$

para $i = 2, 3, \ldots, 9$, que se reduce a

$$w_{i+1} = \frac{1}{24}[1.8w_{i+1} + 27.8w_i - w_{i-1} + 0.2w_{i-2} - 0.192i^2 - 0.192i + 4.736].$$

Para utilizar explícitamente este método, despejamos w_{i+1} y obtenemos

$$w_{i+1} = \frac{1}{22.2}[27.8w_i - w_{i-1} + 0.2w_{i-2} - 0.192i^2 - 0.192i + 4.736],$$

para $i = 2, 3, \ldots, 9$.

Los resultados de la tabla 5.11 se obtuvieron empleando los valores exactos provenientes de $y(t) = (t + 1)^2 - 0.5e^t$ para α, α_1, α_2 y α_3 en el caso de Adams-Bashforth y para α, α_1 y α_2 en el caso de Adams-Moulton. ■

En el ejemplo 4, el método implícito de Adams-Moulton dio mejores resultados que el método explícito de Adams-Bashforth del mismo orden. Aunque generalmente es así, los métodos implícitos tienen la debilidad intrínseca de que primero deben convertir algebraicamente el método en una representación explícita de w_{i+1}. Este procedimiento no siempre es posible, como se advierte al considerar el problema elemental con valor inicial

$$y' = e^y, \quad 0 \le t \le 0.25, \quad y(0) = 1.$$

Tabla 5.11

t_i	Valores exactos	Método Adams-Bashforth w_i	Error	Método Adams-Moulton w_i	Error
0.0	0.5000000				
0.2	0.8292986				
0.4	1.2140877				
0.6	1.6489406			1.6489341	0.0000065
0.8	2.1272295	2.1273124	0.0000828	2.1272136	0.0000160
1.0	2.6408591	2.6410810	0.0002219	2.6408298	0.0000293
1.2	3.1799415	3.1803480	0.0004065	3.1798937	0.0000478
1.4	3.7324000	3.7330601	0.0006601	3.7323270	0.0000731
1.6	4.2834838	4.2844931	0.0010093	4.2833767	0.0001071
1.8	4.8151763	4.8166575	0.0014812	4.8150236	0.0001527
2.0	5.3054720	5.3075838	0.0021119	5.3052587	0.0002132

Dado que $f(t, y) = e^y$, el método de Adams-Moulton de tres pasos tiene

$$w_{i+1} = w_i + \frac{h}{24}[9e^{w_{i+1}} + 19e^{w_i} - 5e^{w_{i-1}} + e^{w_{i-2}}]$$

como su ecuación de diferencias, y esta ecuación no podemos resolverla para w_{i+1}.

Podríamos usar el método de Newton o el de la secante para aproximar w_{i+1}, pero esto complica demasiado el procedimiento.

En la práctica, los métodos multipasos implícitos no se emplean como se explica aquí. Por el contrario, sirven para mejorar las aproximaciones obtenidas con métodos explícitos. La combinación de un método explícito con uno implícito recibe el nombre de **método predictor-corrector.** El método explícito predice una aproximación y el implícito corrige la predicción.

Consideremos el siguiente método de cuarto orden para resolver un problema de valor inicial. El primer paso consiste en calcular los valores iniciales w_0, w_1, w_2 y w_3 con el método de Adams-Bashforth de cuatro pasos. Para ello utilizaremos un método de un paso de cuarto orden, el método de cuarto orden de Runge-Kutta. El siguiente paso consiste en calcular una aproximación $w_4^{(0)}$, $y(t_4)$ usando como predictor el método explícito de Adams-Bashforth:

$$w_4^{(0)} = w_3 + \frac{h}{24}[55f(t_3, w_3) - 59f(t_2, w_2) + 37f(t_1, w_1) - 9f(t_0, w_0)].$$

Esta aproximación mejora mucho si se inserta $w_4^{(0)}$ en el lado derecho del método implícito de Adams-Moulton de tres pasos y aplicándolo como corrector:

$$w_4^{(1)} = w_3 + \frac{h}{24}[9f(t_4, w_4^{(0)}) + 19f(t_3, w_3) - 5f(t_2, w_2) + f(t_1, w_1)].$$

En este procedimiento, la única nueva evaluación de función que se requiere es $f(t_4, w_4^{(0)})$ en la ecuación del corrector. El resto de los valores de f han sido calculados para las aproximaciones anteriores.

Después utilizamos el valor $w_4^{(1)}$ como aproximación a $y(t_4)$, y la técnica que consiste en utilizar como predictor el método de Adams-Bashforth y como corrector el de Adams-Moulton se repite para obtener $w_5^{(0)}$ y $w_5^{(1)}$, las aproximaciones inicial y final de $y(t_5)$, etcétera.

Podemos obtener mejores aproximaciones a $y(t_{i+1})$ iterando la fórmula de Adams-Moulton

$$w_{i+1}^{(k+1)} = w_i + \frac{h}{24}[9f(t_{i+1}, w_{i+1}^{(k)}) + 19f(t_i, w_i) - 5f(t_{i-1}, w_{i-1}) + f(t_{i-2}, w_{i-2})].$$

Sin embargo, $\{w_{i+1}^{(k+1)}\}$ converge a la aproximación dada por la fórmula implícita y no a la solución $y(t_{i+1})$ y suele ser más eficiente usar una reducción del tamaño de paso si se necesita mejorar la exactitud.

El algoritmo 5.4 se basa en el método de Adams-Bashforth de cuarto orden como predictor y en una iteración del método de Adams-Moulton como corrector, con los valores iniciales conseguidos con el método de Runge-Kutta de cuarto orden.

ALGORITMO 5.4

Corrector-predictor de cuarto orden de Adams

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

en $(N + 1)$ números uniformemente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α.

SALIDA aproximación w a y en los $(N + 1)$ valores de t.

Paso 1 Tome $h = (b - a)/N$;
$t_0 = a$;
$w_0 = \alpha$;
SALIDA (t_0, w_0).

Paso 2 Para $i = 1, 2, 3$, haga pasos 3-5.
(*Calcule valores usando el método de Runge-Kutta.*)

Paso 3 Tome $K_1 = hf(t_{i-1}, w_{i-1})$;
$K_2 = hf(t_{i-1} + h/2, w_{i-1} + K_1/2)$;
$K_3 = hf(t_{i-1} + h/2, w_{i-1} + K_2/2)$;
$K_4 = hf(t_{i-1} + h, w_{i-1} + K_3)$;

Paso 4 Tome $w_i = w_{i-1} + (K_1 + 2K_2 + 2K_3 + K_4)/6$;
$t_i = a + ih$.

Paso 5 SALIDA (t_i, w_i).

Paso 6 Para $i = 4, \ldots, N$ haga pasos 7-10.

Paso 7 Tome $t = a + ih$;
$w = w_3 + h[55f(t_3, w_3) - 59f(t_2, w_2) + 37f(t_1, w_1) - 9f(t_0, w_0)]/24$; (*Predice w_i.*)
$w = w_3 + h[9f(t, w) + 19f(t_3, w_3) - 5f(t_2, w_2) + f(t_1, w_1)]/24$. (*Corrige w_i.*)

Paso 8 SALIDA (t, w).

Paso 9 Para $j = 0, 1, 2$
tome $t_j = t_{j+1}$; (*Prepare la siguiente iteración.*)
$w_j = w_{j+1}$.

Paso 10 Tome $t_3 = t$;
$w_3 = w$.

Paso 11 PARAR. ■

EJEMPLO 5 La tabla 5.12 contiene los resultados obtenidos al usar el algoritmo 5.4. Para el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

con $N = 10$. Aquí los resultados son más precisos que en el ejemplo 4, que sólo usaba el corrector (es decir, el método implícito de Adams-Moulton), pero esto no siempre ocurre. ■

Tabla 5.12

t_i	$y_i = y(t_i)$	w_i	Error $\lvert y_i - w_i \rvert$
0.0	0.5000000	0.5000000	0
0.2	0.8292986	0.8292933	0.0000053
0.4	1.2140877	1.2140762	0.0000114
0.6	1.6489406	1.6489220	0.0000186
0.8	2.1272295	2.1272056	0.0000239
1.0	2.6408591	2.6408286	0.0000305
1.2	3.1799415	3.1799026	0.0000389
1.4	3.7324000	3.7323505	0.0000495
1.6	4.2834838	4.2834208	0.0000630
1.8	4.8151763	4.8150964	0.0000799
2.0	5.3054720	5.3053707	0.0001013

Podemos obtener otros métodos multipasos utilizando la integración de los polinomios interpolantes en los intervalos de la forma $[t_j, t_{i+1}]$, con $j \le i - 1$, para obtener una aproximación a $y(t_{i+1})$. Cuando integramos en $[t_{i-3}, t_{i+1}]$ un polinomio interpolante, el resultado será una técnica explícita denominada **método de Milne**,

$$w_{i+1} = w_{i-3} + \frac{4h}{3}[2f(t_i, w_i) - f(t_{i-1}, w_{i-1}) + 2f(t_{i-2}, w_{i-2})],$$

que tiene el error local de truncamiento $\frac{14}{45}h^4 y^{(5)}(\xi_i)$, para alguna $\xi_i \in (t_{i-3}, t_{i+1})$.

En ocasiones, este método se usa como predictor de un **método implícito de Simpson**,

$$w_{i+1} = w_{i-1} + \frac{h}{3}[f(t_{i+1}, w_{i+1}) + 4f(t_i, w_i) + f(t_{i-1}, w_{i-1})],$$

que tiene el error local de truncamiento $-(h^4/90)y^{(5)}(\xi_i)$, para alguna $\xi_i \in (t_{i-1}, t_{i+1})$, y que se obtiene integrando un polinomio interpolante en $[t_{i-1}, t_{i+1}]$.

El error local de truncamiento relacionado con un método predictor-corrector de tipo Milne-Simpson, suele ser menor que el del método de Adams-Bashforth-Moulton; pero se usa poco debido a los problemas de estabilidad, lo que no ocurre con el procedimiento de Adams. En la sección 5.10 se explica más a fondo este problema.

CONJUNTO DE EJERCICIOS 5.6

1. Aplique los métodos de Adams-Bashforth para aproximar las soluciones de los siguientes problemas de valor inicial. En cada caso utilice valores iniciales exactos y después compare los resultados con los valores reales.
   a. $y' = te^{3t} - 2y$, $0 \le t \le 1$, $y(0) = 0$, con $h = 0.2$; solución real $y(t) = \frac{1}{5}te^{3t} - \frac{1}{25}e^{3t} + \frac{1}{25}e^{-2t}$.
   b. $y' = 1 + (t - y)^2$, $2 \le t \le 3$, $y(2) = 1$, con $h = 0.2$; solución real $y(t) = t + \frac{1}{1-t}$.
   c. $y' = 1 + y/t$, $1 \le t \le 2$, $y(1) = 2$, con $h = 0.2$; solución real $y(t) = t \ln t + 2t$.
   d. $y' = \cos 2t + \operatorname{sen} 3t$, $0 \le t \le 1$, $y(0) = 1$, con $h = 0.2$; solución real $y(t) = \frac{1}{2}\operatorname{sen} 2t - \frac{1}{3}\cos 3t + \frac{4}{3}$.
2. Aplique los métodos de Adams-Moulton para aproximar las soluciones de los ejercicios 1(a), 1(c) y 1(d). En cada caso utilice valores iniciales exactos y resuelva explícitamente w_{i+1}. Después, compare los resultados con los valores reales.
3. Aplique los métodos de Adams-Bashforth para aproximar las soluciones de los siguientes problemas de valor inicial. En cada caso utilice los valores iniciales obtenidos con el método de Runge-Kutta de cuarto orden. Después, compare los resultados con los valores reales.
   a. $y' = y/t - (y/t)^2$, $1 \le t \le 2$, $y(1) = 1$, con $h = 0.1$; solución real $y(t) = \frac{t}{1+\ln t}$.
   b. $y' = 1 + y/t + (y/t)^2$, $1 \le t \le 3$, $y(1) = 0$, con $h = 0.2$; solución real $y(t) = t\tan(\ln t)$.
   c. $y' = -(y + 1)(y + 3)$, $0 \le t \le 2$, $y(0) = -2$, con $h = 0.1$; solución real $y(t) = -3 + 2/(1 + e^{-2t})$.
   d. $y' = -5y + 5t^2 + 2t$, $0 \le t \le 1$, $y(0) = 1/3$, con $h = 0.1$; solución real $y(t) = t^2 + \frac{1}{3}e^{-5t}$.
4. Use el algoritmo 5.4 para aproximar las soluciones de los problemas de valor inicial en el ejercicio 1.
5. Use el algoritmo 5.4 para aproximar las soluciones de los problemas de valor inicial en el ejercicio 3.
6. Modifique el algoritmo 5.4 de modo que pueda iterar el corrector en una cantidad de iteraciones p. Repita el ejercicio 5 con $p = 2$, 3 y 4 iteraciones. ¿Cuál elección de p produce la mejor respuesta de cada problema de valor inicial?
7. El problema de valor inicial

$$y' = e^y, \quad 0 \le t \le 0.20, \quad y(0) = 1$$

   tiene la solución

$$y(t) = 1 - \ln(1 - et).$$

   Aplicando a este problema el método de Adams-Moulton de tres pasos es equivalente a encontrar el punto fijo w_{i+1} de

$$g(w) = w_i + \frac{h}{24}\left[9e^{w} + 19e^{w_i} - 5e^{w_{i-1}} + e^{w_{i-2}}\right].$$

a. Con $h = 0.01$, obtenga w_{i+1} mediante la iteración funcional para $i = 2, \ldots, 19$ empleando los valores iniciales exactos w_0, w_1 y w_2. En cada paso utilice w_i para aproximar inicialmente w_{i+1}.

b. ¿Acelerará el método de Newton la convergencia en la iteración funcional?

8. Aplique el método predictor-corrector de Milne-Simpson para aproximar las soluciones de los problemas de valor inicial del ejercicio 3.

9. a. Mediante la forma de Lagrange del polinomio interpolante deduzca la ecuación (5.32).

b. Utilice la forma de la diferencia regresiva de Newton del polinomio interpolante para que deduzca la ecuación (5.34).

10. Derive la ecuación (5.33) con el siguiente método. Use

$$y(t_{i+1}) = y(t_i) + ahf(t_i, y(t_i)) + bhf(t_{i-1}, y(t_{i-1})) + chf(t_{i-2}, y(t_{i-2})).$$

Desarrolle $y(t_{i+1})$, $f(t_{i-2}, y(t_{i-2}))$ y $f(t_{i-1}, y(t_{i-1}))$ en serie de Taylor alrededor de $(t_i, y(t_i))$ e iguale los coeficientes, h, h^2 y h^3 para obtener a, b y c.

11. Deduzca la ecuación (5.36) y su error local de truncamiento, empleando una forma adecuada de un polinomio interpolante.

12. Deduzca el método de Simpson aplicando la regla de Simpson a la integral

$$y(t_{i+1}) - y(t_{i-1}) = \int_{t_{i-1}}^{t_{i+1}} f(t, y(t))\, dt.$$

13. Obtenga el método de Milne aplicando la fórmula abierta de Newton-Cotes (4.29) a la integral

$$y(t_{i+1}) - y(t_{i-3}) = \int_{t_{i-3}}^{t_{i+1}} f(t, y(t))\, dt.$$

14. Verifique los datos de la tabla 5.10.

5.7 Métodos multipasos con tamaño variable de paso

El método de Runge-Kutta-Fehlberg se usa para controlar el error porque cada paso ofrece, con un pequeño costo adicional, *dos* aproximaciones comparables y relacionadas con el error local. Las técnicas predictor-corrector siempre generan *dos* aproximaciones en cada paso, por lo cual son candidatas naturales para adaptar el control del error.

Con el fin de mostrar el procedimiento de control del error, construiremos un método predictor-corrector con tamaño variable de paso, utilizando como predictor el método explícito de Adams-Bashforth de cuatro pasos y como corrector el método implícito de Adams-Moulton de tres pasos.

El método de Adams-Bashforth de cuatro pasos proviene de la relación

$$y(t_{i+1}) = y(t_i) + \frac{h}{24}[55f(t_i, y(t_i)) - 59f(t_{i-1}, y(t_{i-1})) + 37f(t_{i-2}, y(t_{i-2})) - 9f(t_{i-3}, y(t_{i-3}))] + \frac{251}{720}y^{(5)}(\hat{\mu}_i)h^5,$$

para algún $\hat{\mu}_i \in (t_{i-3}, t_{i+1})$. La suposición de que las aproximaciones $w_0, w_1, \ldots, w_i$ son todas exactas, significa que el error de truncamiento de Adams-Bashforth es

$$\frac{y(t_{i+1}) - w_{i+1}^{(0)}}{h} = \frac{251}{720}y^{(5)}(\hat{\mu}_i)h^4. \tag{5.39}$$

Un análisis similar del método de Adams-Moulton de tres pasos, que proviene de

$$y(t_{i+1}) = y(t_i) + \frac{h}{24}[9f(t_{i+1}, y(t_{i+1})) + 19f(t_i, y(t_i)) - 5f(t_{i-1}, y(t_{i-1}))$$

$$+ f(t_{i-2}, y(t_{i-2}))] - \frac{19}{720}y^{(5)}(\tilde{\mu}_i)h^5,$$

para algún $\tilde{\mu}_i \in (t_{i-2}, t_{i+1})$ nos lleva al error local de truncamiento

$$\frac{y(t_{i+1}) - w_{i+1}}{h} = -\frac{19}{720}y^{(5)}(\tilde{\mu}_i)h^4. \tag{5.40}$$

Si queremos avanzar más, debemos suponer que, para un valor pequeño de h,

$$y^{(5)}(\hat{\mu}_i) \approx y^{(5)}(\tilde{\mu}_i).$$

La efectividad del método de control de error depende directamente de esta suposición.

Si restamos la ecuación (5.40) a la ecuación (5.39), tendremos

$$\frac{w_{i+1} - w_{i+1}^{(0)}}{h} = \frac{h^4}{720}[251y^{(5)}(\hat{\mu}_i) + 19y^{(5)}(\tilde{\mu}_i)] \approx \frac{3}{8}h^4y^{(5)}(\tilde{\mu}_i),$$

y, por tanto,

$$y^{(5)}(\tilde{\mu}_i) \approx \frac{8}{3h^5}(w_{i+1} - w_{i+1}^{(0)}). \tag{5.41}$$

Al utilizar este resultado para suprimir el término que contiene $h^4y^{(5)}(\tilde{\mu}_i)$ en (5.40), obtenemos la aproximación al error

$$|\tau_{i+1}(h)| = \frac{|y(t_{i+1}) - w_{i+1}|}{h} \approx \frac{19h^4}{720} \cdot \frac{8}{3h^5}|w_{i+1} - w_{i+1}^{(0)}| = \frac{19|w_{i+1} - w_{i+1}^{(0)}|}{270h}.$$

Supóngase que ahora reconsideramos (5.40) con un nuevo tamaño de paso qh que genera nuevas aproximaciones $\hat{w}_{i+1}^{(0)}$ y $\hat{w}_{i+1}$. El objetivo es escoger q, tal que el error local de truncamiento de (5.40) esté acotado por la tolerancia prescrita ε. Si suponemos que el valor $y^{(5)}(\mu)$ en (5.40) asociado a qh se aproxima también por medio de (5.41), entonces

$$\frac{|y(t_i + qh) - \hat{w}_{i+1}|}{qh} = \frac{19q^4h^4}{720}|y^{(5)}(\mu)| \approx \frac{19q^4h^4}{720}\left[\frac{8}{3h^5}|w_{i+1} - w_{i+1}^{(0)}|\right],$$

y debemos seleccionar q tal que

$$\frac{|y(t_i + qh) - \hat{w}_{i+1}|}{qh} \approx \frac{19q^4}{270}\frac{|w_{i+1} - w_{i+1}^{(0)}|}{h} < \varepsilon.$$

Es decir, seleccionamos q tal que

$$q < \left(\frac{270}{19}\frac{h\varepsilon}{|w_{i+1} - w_{i+1}^{(0)}|}\right)^{1/4} \approx 2\left(\frac{h\varepsilon}{|w_{i+1} - w_{i+1}^{(0)}|}\right)^{1/4}.$$

En este planteamiento hemos hecho varias suposiciones de aproximación, así que en la práctica se selecciona q en forma conservadora, generalmente así

$$q = 1.5\left(\frac{h\varepsilon}{\left|w_{i+1} - w_{i+1}^{(0)}\right|}\right)^{1/4}.$$

Un cambio del tamaño de paso en un método multipasos requiere más evaluaciones de funciones que un método de un paso, porque hay que calcular nuevos valores iniciales uniformemente espaciados. En consecuencia, en la práctica se acostumbra ignorar el cambio de tamaño de paso siempre que el error local de truncamiento se encuentre entre $\varepsilon/10$ y ε, es decir, cuando

$$\frac{\varepsilon}{10} < \left|\tau_{i+1}(h)\right| = \frac{\left|y(t_{i+1}) - w_{i+1}\right|}{h} \approx \frac{19\left|w_{i+1} - w_{i+1}^{(0)}\right|}{270h} < \varepsilon.$$

Además, a q suele asignársele una cota superior para asegurarse de que una sola aproximación de exactitud poco usual no produzca un tamaño de paso demasiado grande. El algoritmo 5.5 incorpora esta medida de seguridad con una cota superior de 4.

Recuerde lo siguiente: como los métodos multipasos requieren tamaños iguales de paso en los valores iniciales, cualquier cambio de tamaño exige recalcular otros valores iniciales en ese punto. En el algoritmo 5.5 esto se hace llamando un subalgoritmo de Runge-Kutta (algoritmo 5.2).

ALGORITMO 5.5

Corrector-predictor con tamaño variable de paso de Adams

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha$$

con un error local de truncamiento dentro los límites de una tolerancia dada:

ENTRADA extremos a, b; condición inicial α; tolerancia TOL; tamaño máximo de paso $hmáx$; tamaño mínimo de paso $hmín$.

SALIDA i, t_i, w_i, h donde en el i-ésimo paso w_i aproxima $y(t_i)$ y se usó el tamaño de paso h, o bien un mensaje de que se rebasó el tamaño mínimo de paso.

Paso 1 Plantee un subalgoritmo para el método Runge-Kutta de cuarto orden al que se asignará el nombre $RK4(h, v_0, x_0, v_1, x_1, v_2, x_2, v_3, x_3)$ que acepta como entrada un tamaño de paso h y valores iniciales $v_0 \approx y(x_0)$ y devuelve $\{(x_j, v_j) \mid j = 1, 2, 3\}$ definido por lo siguiente:
para $j = 1, 2, 3$
tome $K_1 = hf(x_{j-1}, v_{j-1})$;
$K_2 = hf(x_{j-1} + h/2, v_{j-1} + K_1/2)$
$K_3 = hf(x_{j-1} + h/2, v_{j-1} + K_2/2)$
$K_4 = hf(x_{j-1} + h, v_{j-1} + K_3)$
$v_j = v_{j-1} + (K_1 + 2K_2 + 2K_3 + K_4)/6$;
$x_j = x_0 + jh$.

Paso 2 Tome $t_0 = a$;
$w_0 = \alpha$;

$h = hmáx$;
$BAN = 1$; (*BAN será usado para salir del ciclo en el paso 4.*)
$ULT = 0$; (*ULT indica cuándo se calcula el último valor.*)
SALIDA (t_0, w_0).

Paso 3 Llame $RK4(h, w_0, t_0, w_1, t_1, w_2, t_2, w_3, t_3)$;
Tome $NBAN = 1$; (*indica cálculo a partir de RK4*).
$i = 4$;
$t = t_3 + h$.

Paso 4 Mientras ($BAN = 1$) haga pasos 5-20.

Paso 5 Si $WP = w_{i-1} + \frac{h}{24}[55f(t_{i-1}, w_{i-1}) - 59f(t_{i-2}, w_{i-2})$
$+ 37f(t_{i-3}, w_{i-3}) - 9f(t_{i-4}, w_{i-4})]$; (*Predice w_i.*)
$WC = w_{i-1} + \frac{h}{24}[9f(t, WP) + 19f(t_{i-1}, w_{i-1})$
$- 5f(t_{i-2}, w_{i-2}) + f(t_{i-3}, w_{i-3})]$; (*Corrige w_i.*)
$\sigma = 19|WC - WP|/(270h)$.

Paso 6 Si $\sigma \leq TOL$ entonces haga pasos 7-16 (*Resultado aceptado.*)
si no, haga pasos 17-19. (*Resultado rechazado.*)

Paso 7 Tome $w_i = WC$; (*Resultado aceptado.*)
$t_i = t$.

Paso 8 Si $NBAN = 1$ entonces para $j = i - 3, i - 2, i - 1, i$
SALIDA (j, t_j, w_j, h);
(*Los resultados previos también aceptados.*)
si no, SALIDA (i, t_i, w_i, h).
(*Los resultados previos ya aceptados.*)

Paso 9 Si $ULT = 1$ entonces tome $BAN = 0$ (*Siguiente paso es el 20.*)
si no, haga pasos 10-16.

Paso 10 Tome $i = i + 1$;
$NBAN = 0$.

Paso 11 Si $\sigma \leq 0.1\ TOL$ o $t_{i-1} + h > b$ entonces haga pasos 12-16.
(*Aumente h si es más preciso de lo requerido o disminuya h para incluir a b como un punto de red*).

Paso 12 Tome $q = (TOL/(2\sigma))^{1/4}$.

Paso 13 Si $q > 4$ entonces tome $h = 4h$
si no, tome $h = qh$.

Paso 14 Si $h > hmáx$ entonces tome $h = hmáx$.

Paso 15 Si $t_{i-1} + 4h > b$ entonces
tome $h = (b - t_{i-1})/4$;
$ULT = 1$.

Paso 16 Llame $RK4(h, w_{i-1}, t_{i-1}, w_i, t_i, w_{i+1}, t_{i+1}, w_{i+2}, t_{i+2})$;
Tome $NBAN = 1$;
$i = i + 3$. (*Terminada rama verdadera. Siguiente paso es el 20.*)

Paso 17 Tome $q = (TOL/(2\sigma))^{1/4}$. (*Rama falsa desde paso 6: Resultado rechazado.*)

Paso 18 Si $q < 0.1$ entonces tome $h = 0.1h$
si no, tome $h = qh$.

Paso 19 Si $h < hmín$ entonces tome $BAN = 0$;
SALIDA ('*hmín rebasado*')
si no
si $NBAN = 1$ entonces tome $i = i - 3$;
(*Resultados previos también rechazados.*)
Llame $RK4(h, w_{i-1}, t_{i-1}, w_i, t_i, w_{i+1}, t_{i+1}, w_{i+2}, t_{i+2})$;
tome $i = i + 3$;
$NBAN = 1$.

Paso 20 Tome $t = t_{i-1} + h$.

Paso 21 PARAR. ■

EJEMPLO 1 La tabla 5.13 contiene los resultados obtenidos al usar el algoritmo 5.5 para calcular las aproximaciones a la solución del problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

Tabla 5.13

t_i	$y(t_i)$	w_i	h_i	σ_i	$\lvert y(t_i) - w_i \rvert$
0	0.5	0.5			
0.1257017	0.7002323	0.7002318	0.1257017	4.051×10^{-6}	0.0000005
0.2514033	0.9230960	0.9230949	0.1257017	4.051×10^{-6}	0.0000011
0.3771050	1.1673894	1.1673877	0.1257017	4.051×10^{-6}	0.0000017
0.5028066	1.4317502	1.4317480	0.1257017	4.051×10^{-6}	0.0000022
0.6285083	1.7146334	1.7146306	0.1257017	4.610×10^{-6}	0.0000028
0.7542100	2.0142869	2.0142834	0.1257017	5.210×10^{-6}	0.0000035
0.8799116	2.3287244	2.3287200	0.1257017	5.913×10^{-6}	0.0000043
1.0056133	2.6556930	2.6556877	0.1257017	6.706×10^{-6}	0.0000054
1.1313149	2.9926385	2.9926319	0.1257017	7.604×10^{-6}	0.0000066
1.2570166	3.3366642	3.3366562	0.1257017	8.622×10^{-6}	0.0000080
1.3827183	3.6844857	3.6844761	0.1257017	9.777×10^{-6}	0.0000097
1.4857283	3.9697541	3.9697433	0.1030100	7.029×10^{-6}	0.0000108
1.5887383	4.2527830	4.2527711	0.1030100	7.029×10^{-6}	0.0000120
1.6917483	4.5310269	4.5310137	0.1030100	7.029×10^{-6}	0.0000133
1.7947583	4.8016639	4.8016488	0.1030100	7.029×10^{-6}	0.0000151
1.8977683	5.0615660	5.0615488	0.1030100	7.760×10^{-6}	0.0000172
1.9233262	5.1239941	5.1239764	0.0255579	3.918×10^{-8}	0.0000177
1.9488841	5.1854932	5.1854751	0.0255579	3.918×10^{-8}	0.0000181
1.9744421	5.2460056	5.2459870	0.0255579	3.918×10^{-8}	0.0000186
2.0000000	5.3054720	5.3054529	0.0255579	3.918×10^{-8}	0.0000191

que tiene la solución $y(t) = (t + 1)^2 - 0.5e^t$. En la entrada se incluye la tolerancia $TOL = 10^{-5}$, el tamaño máximo de paso $hmáx = 0.25$ y el tamaño mínimo de paso $hmín = 0.01$. ■

CONJUNTO DE EJERCICIOS 5.7

1. Use el algoritmo predictor-corrector con tamaño variable de paso de Adams con una tolerancia $TOL = 10^{-4}$, $hmáx = 0.25$ y $hmín = 0.025$ para aproximar las soluciones de los siguientes problemas de valor inicial. Después, compare los resultados con los valores reales.
   a. $y' = te^{3t} - 2y$, $0 \le t \le 1$, $y(0) = 0$; solución real $y(t) = \frac{1}{5}te^{3t} - \frac{1}{25}e^{3t} + \frac{1}{25}e^{-2t}$.
   b. $y' = 1 + (t - y)^2$, $2 \le t \le 3$, $y(2) = 1$; solución real $y(t) = t + 1/(1 - t)$.
   c. $y' = 1 + y/t$, $1 \le t \le 2$, $y(1) = 2$; solución real $y(t) = t \ln t + 2t$.
   d. $y' = \cos 2t + \operatorname{sen} 3t$, $0 \le t \le 1$, $y(0) = 1$; solución real $y(t) = \frac{1}{2}\operatorname{sen} 2t - \frac{1}{3}\cos 3t + \frac{4}{3}$.
2. Use el algoritmo predictor-corrector con tamaño variable de paso de Adams con una tolerancia $TOL = 10^{-4}$ para aproximar las soluciones de los siguientes problemas de valor inicial:
   a. $y' = (y/t)^2 + y/t$, $1 \le t \le 1.2$, $y(1) = 1$, con $hmáx = 0.05$ y $hmín = 0.01$.
   b. $y' = \operatorname{sen} t + e^{-t}$, $0 \le t \le 1$, $y(0) = 0$, con $hmáx = 0.2$ y $hmín = 0.01$.
   c. $y' = (1/t)(y^2 + y)$, $1 \le t \le 3$, $y(1) = -2$, con $hmáx = 0.4$ y $hmín = 0.01$.
   d. $y' = -ty + 4t/y$, $0 \le t \le 1$, $y(0) = 1$, con $hmáx = 0.2$ y $hmín = 0.01$.
3. Use el algoritmo predictor-corrector con tamaño variable de paso de Adams con una tolerancia $TOL = 10^{-6}$, $hmáx = 0.5$ y $hmín = 0.02$ para aproximar las soluciones de los siguientes problemas de valor inicial. Después, compare los resultados con los valores iniciales.
   a. $y' = y/t - (y/t)^2$, $1 \le t \le 4$, $y(1) = 1$; solución real $y(t) = t/(1 + \ln t)$.
   b. $y' = 1 + y/t + (y/t)^2$, $1 \le t \le 3$, $y(1) = 0$; solución real $y(t) = t \tan(\ln t)$.
   c. $y' = -(y + 1)(y + 3)$, $0 \le t \le 3$, $y(0) = -2$; solución real $y(t) = -3 + 2(1 + e^{-2t})^{-1}$.
   d. $y' = (t + 2t^3)y^3 - ty$, $0 \le t \le 2$, $y(0) = \frac{1}{3}$; solución real $y(t) = (3 + 2t^2 + 6e^{t^2})^{-1/2}$.
4. Construya un algoritmo predictor-corrector con tamaño variable de paso de Adams, tomando como base el método de Adams-Bashforth de cinco pasos y el de Adams-Moulton de cuatro pasos. Repita el ejercicio 3 aplicando este nuevo método.
5. Un circuito eléctrico consiste en un capacitor de capacitancia constante $C = 1.1$ faradios, que está en serie con un resistor de resistencia constante $R_0 = 2.1$ ohms. Se aplica un voltaje $\mathcal{E}(t) = 110 \operatorname{sen} t$ en el tiempo $t = 0$. Cuando el resistor se calienta, la resistencia se transforma en una función de la corriente i,

$$R(t) = R_0 + ki, \qquad \text{donde } k = 0.9,$$

   y la ecuación diferencial de $i(t)$ se convierte en

$$\left(1 + \frac{2k}{R_0}i\right)\frac{di}{dt} + \frac{1}{R_0C}i = \frac{1}{R_0C}\frac{d\mathcal{E}}{dt}.$$

   Calcule $i(2)$, suponiendo que $i(0) = 0$.

5.8 Métodos de extrapolación

En la sección 4.5 utilizamos la extrapolación para aproximar integrales definidas; descubrimos que, al prorratear correctamente las aproximaciones relativamente inexactas del trapecio, podíamos obtener otras que son mucho más precisas. En esta sección aplicaremos la extrapolación para mejorar la exactitud de las aproximaciones a la solución de los problemas de valor inicial. Como explicamos con anterioridad, las aproximaciones originales deben tener un desarrollo del error de forma específica si queremos que el procedimiento sea exitoso.

Para aplicar la extrapolación a la solución de problemas de valor inicial, aplicamos un procedimiento que se basa en el método del punto medio:

$$w_{i+1} = w_{i-1} + 2hf(t_i, w_i), \quad \text{para } i \geq 1. \tag{5.42}$$

Este procedimiento requiere dos valores iniciales, ya que se requieren tanto w_0 como w_1 para poder determinar la primera aproximación al punto medio, w_2. Como de costumbre, usamos la condición inicial con $w_0 = y(a) = \alpha$. Para determinar el segundo valor inicial w_1 aplicamos el método de Euler. Las aproximaciones subsecuentes se obtienen a partir de (5.42). Después de generar una serie de aproximaciones de este tipo que terminan en un valor t, se efectúa una corrección en los extremos que contiene las dos últimas aproximaciones del punto medio. Se consigue así una aproximación $w(t, h)$ a $y(t)$ que tiene la forma

$$y(t) = w(t, h). + \sum_{k=1}^{\infty} \delta_k h^{2k}, \tag{5.43}$$

donde las δ_k son constantes relacionadas con las derivadas de la solución $y(t)$. El punto importante es que las δ_k no dependen del tamaño de paso h. Los detalles de este procedimiento se encuentran en el trabajo de Gragg [Gr].

Para dar un ejemplo del método de extrapolación con que se resuelve

$$y'(t) = f(t, y), \qquad a \leq t \leq b, \qquad y(a) = \alpha,$$

supongamos que tenemos un tamaño de paso h fijo y que queremos aproximar $y(t_1) = y(a + h)$.

En el primer paso de la extrapolación, suponemos que $h_0 = h/2$ y aplicamos el método de Euler con $w_0 = \alpha$ para aproximar $y(a + h_0) = y(a + h/2)$ así

$$w_1 = w_0 + h_0 f(a, w_0).$$

Después aplicamos el método del punto medio con $t_{i-1} = a$ y $t_i = a + h_0 = a + h/2$ para obtener una primera aproximación a $y(a + h) = y(a + 2h_0)$,

$$w_2 = w_0 + 2h_0 f(a + h_0, w_1).$$

Aplicamos la corrección en los extremos para obtener la aproximación final de $y(a + h)$ para el tamaño de paso h_0. Esto nos da la aproximación $O(h_0^2)$ a $y(t_1)$

$$y_{1,1} = \frac{1}{2}[w_2 + w_1 + h_0 f(a + 2h_0, w_2)].$$

Enseguida guardamos la aproximación $y_{1,1}$ y desechamos los resultados intermedios w_1 y w_2.

Para obtener la siguiente aproximación $y_{2,1}$ a $y(t_1)$ usamos $h_1 = h/4$ y el método de Euler con $w_0 = \alpha$ para obtener una aproximación a $y(a + h_1) = y(a + h/4)$ que llamaremos w_1:

$$w_1 = w_0 + h_1 f(a, w_0).$$

Después, calculamos las aproximaciones w_2 a $y(a + 2h_1) = y(a + h/2)$ y w_3 a $y(a + 3h_1) = y(a + 3h/4)$ dadas por

$$w_2 = w_0 + 2h_1 f(a + h_1, w_1) \text{ y } w_3 = w_1 + 2h_1 f(a + 2h_1, w_2).$$

Y luego generamos la aproximación w_4 a $y(a + 4h_1) = y(t_1)$ dada por

$$w_4 = w_2 + 2h_1 f(a + 3h_1, w_3).$$

Enseguida aplicamos la corrección de los extremos a w_3 y a w_4 para obtener la aproximación mejorada $O(h_1^2)$ de $y(t_1)$,

$$y_{2,1} = \frac{1}{2}\,[w_4 + w_3 + h_1 f(a + 4h_1, w_4)].$$

Debido a la forma del error que se muestra en (5.43), las dos aproximaciones a $y(a + h)$ tienen la propiedad de que

$$y(a + h) = y_{1,1} + \delta_1 \left(\frac{h}{2}\right)^2 + \delta_2 \left(\frac{h}{2}\right)^4 + \cdots = y_{1,1} + \delta_1 \frac{h^2}{4} + \delta_2 \frac{h^4}{16} + \cdots,$$

y

$$y(a + h) = y_{2,1} + \delta_1 \left(\frac{h}{4}\right)^2 + \delta_2 \left(\frac{h}{4}\right)^4 + \cdots = y_{2,1} + \delta_1 \frac{h^2}{16} + \delta_2 \frac{h^4}{256} + \cdots.$$

Podemos eliminar la parte $O(h^2)$ de este error de truncamiento, prorrateando adecuadamente estas dos fórmulas. En concreto, si restamos la primera de 4 veces la segunda y dividimos el resultado entre 3, tendremos

$$y(a + h) = y_{2,1} + \frac{1}{3}\,(y_{2,1} - y_{1,1}) - \delta_2 \frac{h^4}{64} + \cdots.$$

Por tanto, la aproximación

$$y_{2,2} = y_{2,1} + \frac{1}{3}\,(y_{2,1} - y_{1,1})$$

tiene un error de orden $O(h^4)$.

Del mismo modo, tomamos $h_2 = h/6$ y aplicamos una vez el método de Euler y luego cinco veces el del punto medio. Después usamos de la corrección de los extremos para determinar la aproximación h^2, $y_{3,1}$ a $y(a + h)$. Podemos prorratear esta aproximación con $y_{2,1}$ y obtener así una segunda aproximación $O(h^4)$ que denotamos $y_{3,2}$. Luego prorrateamos $y_{3,2}$ y $y_{2,2}$, para suprimir los términos de error $O(h^4)$ y producir una aproximación del orden $O(h^6)$. Al continuar el proceso generamos las fórmulas de orden superior.

La única diferencia significativa entre la extrapolación realizada aquí y la utilizada en la integración de Romberg en la sección 4.5, radica en la forma de escoger las subdivisiones. En la integración de Romberg hay una fórmula adecuada para representar las aproxi-

maciones efectuadas mediante la regla del trapecio que utiliza divisiones consecutivas del tamaño de paso por medio de los enteros 1, 2, 4, 8, 16, 32, 64, ... Este procedimiento permite efectuar fácilmente el prorrateo.

No tenemos un medio para producir fácilmente aproximaciones refinadas en los problemas de valor inicial; por ello, seleccionamos las divisiones del método de extrapolación que reduzcan al mínimo la cantidad de evaluaciones de funciones requeridas. El prorrateo que se produce con esta elección de subdivisión, y que se incluye en la tabla 5.14, no es elemental, pero con esa salvedad es el mismo que se emplea en la integración de Romberg.

Tabla 5.14

$y_{1,1} = w(t, h_0)$		
$y_{2,1} = w(t, h_1)$	$y_{2,2} = y_{2,1} + \dfrac{h_1^2}{h_0^2 - h_1^2}(y_{2,1} - y_{1,1})$	
$y_{3,1} = w(t, h_2)$	$y_{3,2} = y_{3,1} + \dfrac{h_2^2}{h_1^2 - h_2^2}(y_{3,1} - y_{2,1})$	$y_{3,3} = y_{3,2} + \dfrac{h_2^2}{h_0^2 - h_2^2}(y_{3,2} - y_{2,2})$

En el algoritmo 5.6 se usa el método de extrapolación con la sucesión de enteros

$$q_0 = 2,\ q_1 = 4,\ q_2 = 6,\ q_3 = 8,\ q_4 = 12,\ q_5 = 16,\ q_6 = 24 \quad \text{y} \quad q_7 = 32.$$

Seleccionamos un tamaño de paso h básico y el método avanza utilizando $h_i = h/q_i$ para cada $i = 0, \ldots, 7$, para aproximar $y(t + h)$. El error se controla al exigir que las aproximaciones $y_{1,1}, y_{2,2}, \ldots,$ se calculen hasta que $|y_{i,i} - y_{i-1,i-1}|$ sea menor que una tolerancia determinada. Si esta última no se logra mediante $i = 8$, entonces reducimos h y repetimos el proceso. Especificamos los valores máximo y mínimo de h, *hmín* y *hmáx*, respectivamente, de modo que garanticen el control del método. Si comprobamos que $y_{i,i}$ es aceptable, entonces transformamos w_1, en $y_{i,i}$ y reanudamos los cálculos para determinar w_2 que aproximará $y(t_2) = y(a + 2h)$. El proceso se repite hasta encontrar la aproximación w_N a $y(b)$.

ALGORITMO 5.6

Extrapolación

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

con el error local de truncamiento dentro de una tolerancia determinada:

ENTRADA extremos, a, b; condición inicial α; tolerancia *TOL*; tamaño máximo de paso *hmáx*; tamaño mínimo de paso *hmín*.

SALIDA T, W, h, donde W aproxima $y(t)$ y se usa el tamaño de paso h o bien un mensaje de que se ha rebasado el tamaño de paso.

Paso 1 Inicialice el arreglo $NK = (2, 4, 6, 8, 12, 16, 24, 32)$.

Paso 2 Tome $TO = a$;
$WO = \alpha$;
$h = hmáx$;
$BAND = 1$. (*Se usa BAND para salir del ciclo en el paso 4.*)

Paso 3 Para $i = 1, 2, \ldots, 7$
para $j = 1, \ldots, i$
tome $Q_{i,j} = (NK_{i+1}/NK_j)^2$. (*Nota:* $Q_{i,j} = h_j^2/h_{i+1}^2$.)

Paso 4 Mientras ($BAND = 1$) haga pasos 5-20.

Paso 5 Tome $k = 1$;
$NBAND = 0$. (*Cuando se obtiene la exactitud deseada, NBAND se toma como* 1.)

Paso 6 Mientras ($k \leq 8$ y $NBAND = 0$) haga pasos 7-14.

Paso 7 Tome $HK = h/NK_k$;
$T = TO$;
$W2 = WO$;
$W3 = W2 + HK \cdot f(T, W_2)$; (*Primer paso de Euler.*)
$T = TO + HK$.

Paso 8 Para $j = 1, \ldots, NK_k - 1$
haga $W1 = W2$;
$W2 = W3$;
$W3 = W1 + 2HK \cdot f(T, W2)$; (*Método del punto medio.*)
$T = TO + (j + 1) \cdot HK$.

Paso 9 Haga $y_k = [W3 + W2 + HK \cdot f(T, W3)]/2$.
(*Correción del punto final para calcular* $y_{k,1}$.)

Paso 10 Si $k \geq 2$ entonces haga pasos 11-13.
(*Nota:* $y_{k-1} \equiv y_{k-1,1}, y_{k-2} \equiv y_{k-2,2}, \ldots, y_1 \equiv y_{k-1,k-1}$ *ya que sólo se guarda el renglón anterior de la tabla.*)

Paso 11 Tome $j = k$;
$v = y_1$. (*Guarde* $y_{k-1,k-1}$.)

Paso 12 Mientras ($j \geq 2$) haga

$$\text{tome } y_{j-1} = y_j + \frac{y_j - y_{j-1}}{Q_{k-1,j-1} - 1};$$

(*Extrapolación para calcular* $y_{j-1} \equiv y_{k,k-j+2}$.)

$$\left(\textit{Nota:} \quad y_{j-1} = \frac{h_{j-1}^2 y_j - h_k^2 y_{j-1}}{h_{j-1}^2 - h_k^2}.\right)$$

$j = j - 1$.

Paso 13 Si $|y_1 - v| \leq TOL$ entonces tome $NBAND = 1$.
(y_1 *se acepta como la nueva* w.)

Paso 14 Tome $k = k + 1$.

Paso 15 Tome $k = k - 1$.

Paso 16 Si $NBAND = 0$ entonces haga pasos 17 y 18 (*Resultado rechazado.*)
si no, haga pasos 19 y 20. (*Resultado aceptado.*)

Paso 17 Tome $h = h/2$. (*Nuevo valor para w rechazado, disminuya h.*)

Paso 18 Si $h < hmín$ entonces
SALIDA ('*hmín* rebasado')
Tome $BAND = 0$.
(*Rama verdadera terminada, siguiente paso retorno a paso 4.*)

Paso 19 Tome $WO = y_1$; (*Nuevo valor de w aceptado.*)
$TO = TO + h$;
SALIDA (TO, WO, h).

Paso 20 Si $TO \geq b$ entonces tome $BAND = 0$
(*Procedimiento terminado exitosamente.*)
si no, si $TO + h > b$ entonces tome $h = b - TO$
(*Termina en $t = b$.*)
si no, si ($k \leq 3$ y $h < 0.5(hmáx)$) entonces tome $h = 2h$.
(*Aumente el tamaño de paso si es posible.*)

Paso 21 PARAR ■

EJEMPLO 1 Consideremos el problema de valor inicial

$$y' = y - t^2 + 1, \quad 0 \leq t \leq 2, \quad y(0) = 0.5,$$

que tiene la solución $y(t) = (t + 1)^2 - 0.5e^t$. Aplicaremos el algoritmo de extrapolación a este problema con $h = 0.25$, $TOL = 10^{-10}$, $hmáx = 0.25$ y con $hmín = 0.01$. La tabla 5.15 se obtiene en el cálculo de w_1. ■

Los cálculos se interrumpen con $w_1 = y_{5,5}$ porque $|y_{5,5} - y_{4,4}| \leq 10^{-10}$ y $y_{5,5}$ se acepta como aproximación de $y(t_1) = y(0.25)$. En la tabla 5.16 viene la serie completa de aproximaciones con las cifras decimales indicadas.

Tabla 5.15

$y_{1,1} = 0.9187011719$				
$y_{2,1} = 0.9200379848$	$y_{2,2} = 0.9204835892$			
$y_{3,1} = 0.9202873689$	$y_{3,2} = 0.9204868761$	$y_{3,3} = 0.9204872870$		
$y_{4,1} = 0.9203747896$	$y_{4,2} = 0.9204871876$	$y_{4,3} = 0.9204872914$	$y_{4,4} = 0.9204872917$	
$y_{5,1} = 0.9204372763$	$y_{5,2} = 0.9204872656$	$y_{5,3} = 0.9204872916$	$y_{5,4} = 0.9204872917$	$y_{5,5} = 0.9204872917$

Tabla 5.16

t_i	$y_i = y(t_i)$	w_i	h_i	k
0.25	0.9204872917	0.9204872917	0.25	5
0.50	1.4256393646	1.4256393646	0.25	5
0.75	2.0039999917	2.0039999917	0.25	5
1.00	2.6408590858	2.6408590858	0.25	5
1.25	3.3173285213	3.3173285212	0.25	4
1.50	4.0091554648	4.0091554648	0.25	3
1.75	4.6851986620	4.6851986619	0.25	3
2.00	5.3054719505	5.3054719505	0.25	3

La demostración de que el método presentado en el algoritmo 5.6 converge, incluye resultados tomados de la teoría de la sumabilidad y se encuentra en el trabajo original de Gragg [Gr]. Existen otros procedimientos de extrapolación, algunos de los cuales utilizan los métodos del tamaño de paso variable. Otros más que se basan en el proceso de extrapolación se explican en los trabajos de Bulirsch y Stoer [BS1], [BS2], [BS3] o en el libro de Stetter [Stet]. Los que emplean Bulirsch y Stoer incluyen la interpolación con funciones racionales en vez de la interpolación polinómica que se usa en el procedimiento de Gragg.

CONJUNTO DE EJERCICIOS 5.8

1. Use el algoritmo de extrapolación con la tolerancia $TOL = 10^{-4}$, $hmáx = 0.25$ y $hmín = 0.05$ para aproximar las soluciones de los siguientes problemas de valor inicial. Compare los resultados con los valores reales.
   a. $y' = te^{3t} - 2y$, $0 \le t \le 1$, $y(0) = 0$; solución real $y(t) = \frac{1}{5}te^{3t} - \frac{1}{25}e^{3t} + \frac{1}{25}e^{-2t}$.
   b. $y' = 1 + (t - y)^2$, $2 \le t \le 3$, $y(2) = 1$; solución real $y(t) = t + 1/(1 - t)$.
   c. $y' = 1 + y/t$, $1 \le t \le 2$, $y(1) = 2$; solución real $y(t) = t \ln t + 2t$.
   d. $y' = \cos 2t + \operatorname{sen} 3t$, $0 \le t \le 1$, $y(0) = 1$; solución real $y(t) = \frac{1}{2}\operatorname{sen} 2t - \frac{1}{3}\cos 3t + \frac{4}{3}$.
2. Use el algoritmo de extrapolación con $TOL = 10^{-4}$ para aproximar las soluciones de los siguientes problemas de valor inicial:
   a. $y' = (y/t)^2 + y/t$, $1 \le t \le 1.2$, $y(1) = 1$, con $hmáx = 0.05$ y $hmín = 0.02$.
   b. $y' = \operatorname{sen} t + e^{-t}$, $0 \le t \le 1$, $y(0) = 0$, con $hmáx = 0.25$ y $hmín = 0.02$.
   c. $y' = (1/t)(y^2 + y)$, $1 \le t \le 3$, $y(1) = -2$, con $hmáx = 0.5$ y $hmín = 0.02$.
   d. $y' = -ty + 4t/y$, $0 \le t \le 1$, $y(0) = 1$, con $hmáx = 0.25$ y $hmín = 0.02$.
3. Use el algoritmo de extrapolación con la tolerancia $TOL = 10^{-6}$, $hmáx = 0.5$ y $hmín = 0.05$ para aproximar las soluciones de los siguientes problemas de valor inicial. Después, compare los resultados con los valores reales.
   a. $y' = y/t - (y/t)^2$, $1 \le t \le 4$, $y(1) = 1$; solución real $y(t) = t/(1 + \ln t)$.
   b. $y' = 1 + y/t + (y/t)^2$, $1 \le t \le 3$, $y(1) = 0$; solución real $y(t) = t \tan(\ln t)$.
   c. $y' = -(y + 1)(y + 3)$, $0 \le t \le 3$, $y(0) = -2$; solución real $y(t) = -3 + 2(1 + e^{-2t})^{-1}$.
   d. $y' = (t + 2t^3)y^3 - ty$, $0 \le t \le 2$, $y(0) = \frac{1}{3}$; solución real $y(t) = (3 + 2t^2 + 6e^{t^2})^{-1/2}$.
4. Sea $P(t)$ el número de individuos de una población en el tiempo t, medido en años. Si la tasa de natalidad promedio b es constante y la tasa de mortalidad promedio d es proporcional al tamaño de la población (debido a la sobrepoblación), entonces la tasa de crecimiento demográfico estará dada por la **ecuación logística**

$$\frac{d\,P(t)}{dt} = b\,P(t) - k[P(t)]^2,$$

   donde $d = kP(t)$. Suponga que $P(0) = 50,976$, $b = 2.9 \times 10^{-2}$ y que $k = 1.4 \times 10^{-7}$. Calcule la población después de 5 años.

5.9 Ecuaciones de orden superior y sistemas de ecuaciones diferenciales

En esta sección presentamos una introducción a la solución numérica de las ecuaciones diferenciales de orden superior, sujetas a condiciones iniciales. Los métodos que se explican son exclusivamente los que transforman una ecuación de orden superior en un sistema de ecuaciones diferenciales de primer orden. Antes de describir el procedimiento de transformación, conviene hacer algunos comentarios sobre los sistemas que contienen ecuaciones diferenciales de primer grado.

Un **sistema de orden** m de problemas de valor inicial de primer orden pueden expresarse como

$$\begin{aligned} \frac{du_1}{dt} &= f_1(t, u_1, u_2, \ldots, u_m), \\ \frac{du_2}{dt} &= f_2(t, u_1, u_2, \ldots, u_m), \\ &\vdots \\ \frac{du_m}{dt} &= f_m(t, u_1, u_2, \ldots, u_m), \end{aligned} \tag{5.44}$$

para a $\leq t \leq b$ con las condiciones iniciales

$$u_1(a) = \alpha_1, \quad u_2(a) = \alpha_2, \ldots, u_m(a) = \alpha_m. \tag{5.45}$$

La finalidad es encontrar m funciones $u_1, u_2, \ldots, u_m$ que satisfagan el sistema de ecuaciones diferenciales y también todas las condiciones iniciales.

Para explicar la existencia y la unicidad de las soluciones de los sistemas de ecuaciones, es necesario extender la definición de la condición de Lipschitz a las funciones de algunas variables.

Definición 5.16 Se dice que la función $f(t, y_1, \ldots, y_m)$ definida en el conjunto

$$D = \{(t, u_1, \ldots, u_m) \mid a \leq t \leq b, -\infty < u_i < \infty, \text{ para cada } i = 1, 2, \ldots, m\}$$

satisface una **condición de Lipschitz** sobre D en las variables $u_1, u_2, \ldots, u_m$ si existe una constante $L > 0$ con la propiedad de que

$$|f(t, u_1, \ldots, u_m) - f(t, z_1, \ldots, z_m)| \leq L \sum_{j=1}^{m} |u_j - z_j|, \tag{5.46}$$

para todo $(t, u_1, \ldots, u_m)$ y $(t, z_1, \ldots, z_m)$ en D. ■

Al utilizar el teorema del valor medio, podemos demostrar que, si f y sus primeras derivadas parciales son continuas en D y que si

$$\left| \frac{\partial f(t, u_1, \ldots, u_m)}{\partial u_i} \right| \leq L,$$

para cada $i = 1, 2, \ldots, m$ y para todo $(t, u_1, \ldots, u_m)$ en D, entonces f satisfará una condición de Lipschitz en D con la constante L de Lipschitz (véase [BiR, p. 141]). Enseguida se incluye un teorema básico de existencia y unicidad. Su demostración se puede encontrar en [BiR, pp. 152-154].

Teorema 5.17 Supongamos que

$$D = \{(t, u_1, u_2, \ldots, u_m) \mid a \le t \le b, -\infty < u_i < \infty, \text{ para cada } i = 1, 2, \ldots, m\},$$

y que $f_i(t, u_1, \ldots, u_m)$ para cada $i = 1, 2, \ldots, m$ es continua en D y que satisface allí una condición de Lipschitz. El sistema de ecuaciones diferenciales de primer orden (5.44), sujeto a las condiciones iniciales (5.45), tiene una solución única $u_1(t), \ldots, u_m(t)$ para $a \le t \le b$. ■

Los métodos con que se resuelven las ecuaciones diferenciales de primer orden son generalizaciones de los métodos para una ecuación de primer orden, explicados anteriormente en este capítulo. Por ejemplo, el método clásico de Runge-Kutta de cuarto orden dado por

$$w_0 = \alpha,$$
$$k_1 = h f(t_i, w_i),$$
$$k_2 = h f(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1),$$
$$k_3 = h f\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right),$$
$$k_4 = h f(t_{i+1}, w_i + k_3),$$

y por

$$w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad \text{para cada } i = 0, 1, \ldots, N-1,$$

con que se resuelve el problema de valor inicial de primer orden

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

se generaliza como sigue.

Seleccionemos un entero $N > 0$ y usemos $h = (b - a)/N$. La partición del intervalo $[a, b]$ en N subintervalos con los puntos de red

$$t_j = a + jh \quad \text{para cada} \quad j = 0, 1, \ldots, N.$$

Use la notación w_{ij} para denotar una aproximación a $u_i(t_j)$ para cada $j = 0, 1, \ldots, N$ y para cada $i = 1, 2, \ldots, m$. Es decir, w_{ij} aproxima la i-ésima solución $u_i(t)$ de (5.44) en el j-ésimo punto de red t_j. En las condiciones iniciales, use (véase Fig. 5.5)

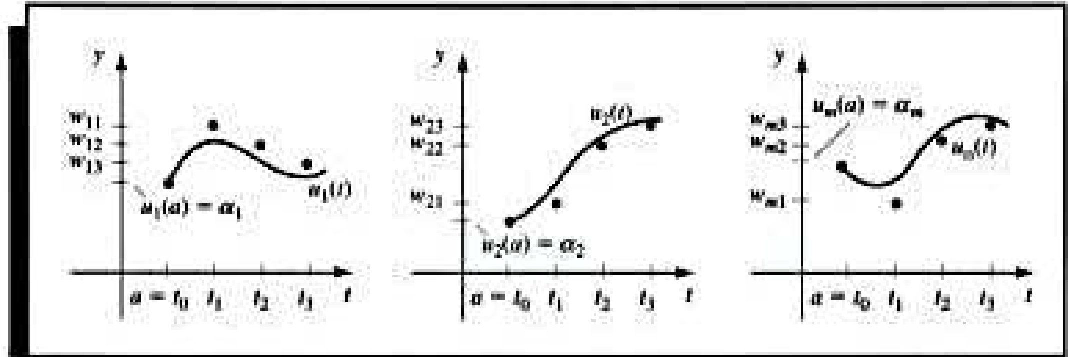
$$w_{1,0} = \alpha_1, \quad w_{2,0} = \alpha_2, \ldots, \quad w_{m,0} = \alpha_m. \tag{5.47}$$

Suponga que se calcularon los valores $w_{1,j}, w_{2,j}, \ldots, w_{m,j}$. Obtenemos $w_{1,j+1}, w_{2,j+1}, \ldots, w_{m,j+1}$ calculando primero

$$k_{1,i} = hf_i(t_j, w_{1,j}, w_{2,j}, \ldots, w_{m,j}), \quad \text{para cada } i = 1, 2, \ldots, m; \tag{5.48}$$

$$k_{2,i} = hf_i\left(t_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{1,1}, w_{2,j} + \frac{1}{2}k_{1,2}, \ldots, w_{m,j} + \frac{1}{2}k_{1,m}\right), \tag{5.49}$$

Figura 5.5



para cada $i = 1, 2, \ldots, m$;

$$k_{3,i} = hf_i\left(t_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{2,1}, w_{2,j} + \frac{1}{2}k_{2,2}, \ldots, w_{m,j} + \frac{1}{2}k_{2,m}\right), \tag{5.50}$$

para cada $i = 1, 2, \ldots, m$;

$$k_{4,i} = hf_i(t_j + h, w_{1,j} + k_{3,1}, w_{2,j} + k_{3,2}, \ldots, w_{m,j} + k_{3,m}), \tag{5.51}$$

para cada $i = 1, 2, \ldots, m$; y entonces

$$w_{i,j+1} = w_{i,j} + \frac{1}{6}(k_{1,i} + 2k_{2,i} + 2k_{3,i} + k_{4,i}), \tag{5.52}$$

para cada $i = 1, 2, \ldots, m$. Nótese que antes de poder determinar cualquiera de los términos de la forma $k_{2,i}$ deben calcularse todos los valores $k_{1,1}, k_{1,2}, \ldots, k_{1,m}$. En general, cada $k_{l,1}, k_{l,2}, \ldots, k_{l,m}$ debe calcularse antes de cualquiera de las expresiones $k_{l+1,i}$. En el algoritmo 5.7 se ejecuta el método de Runge-Kutta de cuarto orden para los sistemas de problemas de valor inicial.

ALGORITMO 5.7

Método de Runge-Kutta para los sistemas de ecuaciones diferenciales

Para aproximar la solución del sistema de m-ésimo orden de los problemas de valor inicial de primer orden

$$u_j' = f_j(t, u_1, u_2, \ldots, u_m), \quad a \le t \le b, \quad \text{con} \quad u_j(a) = \alpha_j,$$

para $j = 1, 2, \ldots, m$ en $(N + 1)$ números uniformemente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; número de ecuaciones m; entero N; condiciones iniciales $\alpha_1, \ldots, \alpha_m$.

SALIDA aproximaciones w_j a $u_j(t)$ en los $(N + 1)$ valores de t.

Paso 1 Tome $h = (b - a)/N$;
$t = a$.

Paso 2 Para $j = 1, 2, \ldots, m$ tome $w_j = \alpha_j$.

Paso 3 SALIDA $(t, w_1, w_2, \ldots, w_m)$.

Paso 4 Para $i = 1, 2, \ldots, N$ haga los pasos 5-11.

Paso 5 Para $j = 1, 2, \ldots, m$ tome
$k_{1,j} = hf_j(t, w_1, w_2, \ldots, w_m)$.

Paso 6 Para $j = 1, 2, \ldots, m$ tome
$k_{2,j} = hf_j\left(t + \frac{h}{2}, w_1 + \frac{1}{2}k_{1,1}, w_2 + \frac{1}{2}k_{1,2}, \ldots, w_m + \frac{1}{2}k_{1,m}\right)$.

Paso 7 Para $j = 1, 2, \ldots, m$ tome
$k_{3,j} = hf_j\left(t + \frac{h}{2}, w_1 + \frac{1}{2}k_{2,1}, w_2 + \frac{1}{2}k_{2,2}, \ldots, w_m + \frac{1}{2}k_{2,m}\right)$.

Paso 8 Para $j = 1, 2, \ldots, m$ tome
$k_{4,j} = hf_j(t + h, w_1 + k_{3,1}, w_2 + k_{3,2}, \ldots, w_m + k_{3,m})$.

Paso 9 Para $j = 1, 2, \ldots, m$ tome
$w_j = w_j + (k_{1,j} + 2k_{2,j} + 2k_{3,j} + k_{4,j})/6$.

Paso 10 Tome $t = a + ih$.

Paso 11 SALIDA $(t, w_1, w_2, \ldots, w_m)$.

Paso 12 PARAR. ■

EJEMPLO 1 La ley de Kirchhoff establece que la suma de todos los cambios instantáneos de voltaje alrededor de un circuito cerrado es cero. Esta ley implica que en un circuito cerrado que contenga una resistencia de R ohms, una capacitancia de C faradios, una inductancia de L henrios y una fuente de voltaje de $E(t)$ voltios, la corriente $I(t)$ satisface la ecuación

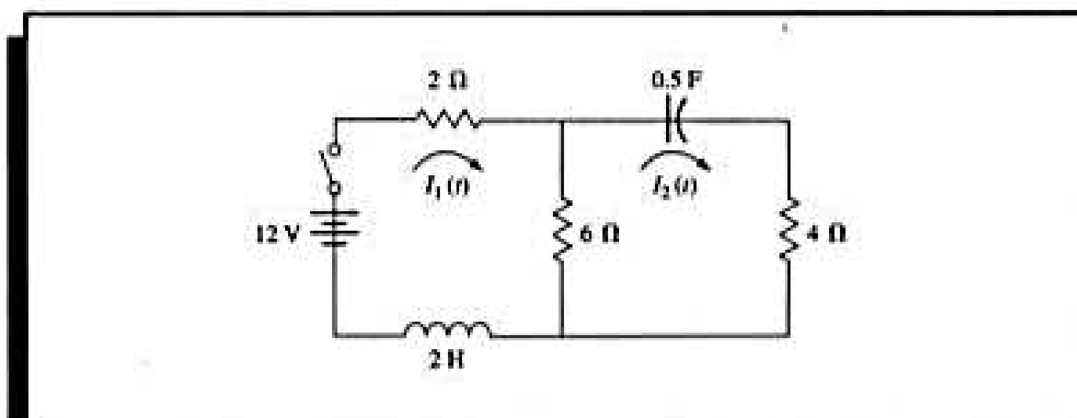
$$LI'(t) + RI(t) + \frac{1}{C}\int I(t)\,dt = E(t).$$

Las corrientes $I_1(t)$ e $I_2(t)$ en los ciclos izquierdo y derecho, respectivamente, del circuito que se muestra en la figura 5.6 son soluciones del sistema de ecuaciones

$$2I_1(t) + 6[I_1(t) - I_2(t)] + 2I_1'(t) = 12,$$

$$\frac{1}{0.5}\int I_2(t)\,dt + 4I_2(t) + 6[I_2(t) - I_1(t)] = 0.$$

Figura 5.6



Supongamos que el interruptor del circuito se encuentra cerrado en el instante $t = 0$. Entonces $I_1(0) = 0$ e $I_2(0) = 0$. Se resuelve para $I_1'(t)$, al diferenciar la segunda ecuación y al sustituir la primera en el resultado, se obtiene el sistema

$$I_1' = f_1(t, I_1, I_2) = -4I_1 + 3I_2 + 6, \quad I_1(0) = 0,$$
$$I_2' = f_2(t, I_1, I_2) = 0.6I_1' - 0.2I_2 = -2.4I_1 + 1.6I_2 + 3.6, \quad I_2(0) = 0.$$

La solución exacta del sistema es

$$I_1(t) = -3.375e^{-2t} + 1.875e^{-0.4t} + 1.5,$$
$$I_2(t) = -2.25e^{-2t} + 2.25e^{-0.4t}.$$

Aplicaremos el método de Runge-Kutta de cuarto orden con $h = 0.1$ a este sistema. Dado que $w_{1,0} = I_1(0) = 0$ y $w_{2,0} = I_2(0) = 0$,

$$k_{1,1} = hf_1(t_0, w_{1,0}, w_{2,0}) = 0.1\, f_1(0, 0, 0) = 0.1[-4(0) + 3(0) + 6] = 0.6,$$
$$k_{1,2} = hf_2(t_0, w_{1,0}, w_{2,0}) = 0.1\, f_2(0, 0, 0) = 0.1[-2.4(0) + 1.6(0) + 3.6] = 0.36,$$
$$k_{2,1} = hf_1\left(t_0 + \frac{1}{2}h, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right) = 0.1\, f_1(0.05, 0.3, 0.18)$$
$$= 0.1[-4(0.3) + 3(0.18) + 6] = 0.534,$$
$$k_{2,2} = hf_2\left(t_0 + \frac{1}{2}h, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right) = 0.1\, f_2(0.05, 0.3, 0.18)$$
$$= 0.1[-2.4(0.3) + 1.6(0.18) + 3.6] = 0.3168.$$

Al generar los valores restantes en una forma semejante, se obtiene

$$k_{3,1} = (0.1)\, f_1(0.05, 0.267, 0.1584) = 0.54072,$$
$$k_{3,2} = (0.1)\, f_2(0.05, 0.267, 0.1584) = 0.321264,$$
$$k_{4,1} = (0.1)\, f_1(0.1, 0.54072, 0.321264) = 0.4800912,$$

y

$$k_{4,2} = (0.1)\, f_2(0.1, 0.54072, 0.321264) = 0.28162944.$$

En consecuencia,

$$I_1(0.1) \approx w_{1,1} = w_{1,0} + \frac{1}{6}\,(k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1})$$
$$= 0 + \frac{1}{6}\,[0.6 + 2(0.534) + 2(0.54072) + 0.4800912] = 0.5382552$$

y

$$I_2(0.1) \approx w_{2,1} = w_{2,0} + \frac{1}{6}(k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2}) = 0.3196263.$$

El resto de los valores de la tabla 5.17 se generan de manera parecida. ■

Tabla 5.17

t_j	$w_{1,j}$	$w_{2,j}$	$\lvert u_1(t_j) - w_{1,j}\rvert$	$\lvert u_2(t_j) - w_{2,j}\rvert$
0.0	0	0	0	0
0.1	0.5382550	0.3196263	0.8285×10^{-5}	0.5803×10^{-5}
0.2	0.9684983	0.5687817	0.1514×10^{-4}	0.9596×10^{-5}
0.3	1.310717	0.7607328	0.1907×10^{-4}	0.1216×10^{-4}
0.4	1.581263	0.9063208	0.2098×10^{-4}	0.1311×10^{-4}
0.5	1.793505	1.014402	0.2193×10^{-4}	0.1240×10^{-4}

El comando de Maple `dsolve` puede usarse para resolver sistemas de ecuaciones diferenciales de primer orden. El sistema del ejemplo 1 se define con

```
>sys2:=D(u1)(t)=-4*u1(t)+3*u2(t)+6,D(u2)(t)=-2.4*u1(t)+1.6*u2(t)+3.6;
```

y la condición inicial con

```
>init2:=u1(0)=0,u2(0)=0;
```

El sistema se resuelve aplicando el comando

```
>sol2:=dsolve({sys2,init2},{u1(t),u2(t)});
```

para obtener

$$\text{sol2} := \left\{u1(t) = \frac{3}{2} - \frac{27}{8}e^{(-2t)} + \frac{15}{8}e^{(-2/5t)},\ u2(t) = -\frac{9}{4}e^{(-2t)} + \frac{9}{4}e^{(-2/5t)}\right\}$$

Si queremos aislar la solución en forma de función, utilizaremos

```
>r1:=rhs(sol2[2]);
```

$$r1 := \frac{3}{2} - \frac{27}{8}e^{(-2t)} + \frac{15}{8}e^{(-2/5t)}$$

y

```
>r2:=rhs(sol2[1]);
```

que nos da una respuesta semejante.

Si queremos evaluar $u_1(0.5)$ y $u_2(0.5)$ usamos

```
>evalf(subs(t=0.5,r1));evalf(subs(t=0.5,r2));
```

para obtener 1.793527048 y 1.014415451.

El comando `dsolve` fallará si no se puede obtener una solución explícita. En tal caso podemos usar la opción numérica en `dsolve`, la cual aplica el método de Runge-Kutta-Fehlberg. Por ejemplo,

```
>g:=dsolve({sys2,init2},{u1(t),u2(t)},numeric);
```

regresa el procedimiento

$$g := \text{proc}(rkf45_x) \ldots \text{fin proc}$$

Para aproximar la solución con $t = 0.5$, introducimos

```
>g(0.5);
```

para obtener

$$[t = .5, u2(t) = 1.0144154547029176l, u1(t) = 1.7935270524376658 6]$$

Muchos problemas importantes de la física (por ejemplo, circuitos eléctricos y sistemas con vibración) implican problemas de valor inicial cuyas ecuaciones tienen orden mayor que uno. No se requieren nuevas técnicas para resolver estos problemas; reetiquetando las variables se puede reducir una ecuación diferencial de orden superior a un sistema de ecuaciones diferenciales de primer orden y luego aplicar uno de los métodos ya analizados. Un problema general de valor inicial del m-ésimo orden

$$y^{(m)}(t) = f(t, y, y', \ldots, y^{(m-1)}), \quad a \le t \le b,$$

con las condiciones iniciales $y(a) = \alpha_1, y'(a) = \alpha_2, \ldots, y^{(m-1)}(a) = \alpha_m$ puede convertirse en un sistema de ecuaciones de la forma (5.44) y (5.45).

Sean $u_1(t) = y(t)$ $u_2(t), = y'(t), \ldots,$ y $u_m(t) = y^{(m-1)}(t)$. Con esto se produce el sistema de primer orden

$$\frac{du_1}{dt} = \frac{dy}{dt} = u_2,$$
$$\frac{du_2}{dt} = \frac{dy'}{dt} = u_3,$$
$$\vdots$$
$$\frac{du_{m-1}}{dt} = \frac{dy^{(m-2)}}{dt} = u_m,$$

y

$$\frac{du_m}{dt} = \frac{dy^{(m-1)}}{dt} = y^{(m)} = f(t, y, y', \ldots, y^{(m-1)}) = f(t, u_1, u_2, \ldots, u_m),$$

con las condiciones iniciales

$$u_1(a) = y(a) = \alpha_1, \quad u_2(a) = y'(a) = \alpha_2, \ldots, \quad u_m(a) = y^{(m-1)}(a) = \alpha_m.$$

EJEMPLO 2 Consideremos el problema de valor inicial de segundo orden

$$y'' - 2y' + 2y = e^{2t} \operatorname{sen} t, \quad \text{para } 0 \le t \le 1, \quad \text{con } y(0) = -0.4, y'(0) = -0.6.$$

Con $u_1(t) = y(t)$ y $u_2(t), = y'(t)$, transformamos esta ecuación en el sistema

$$u_1'(t) = u_2(t),$$
$$u_2'(t) = e^{2t} \operatorname{sen} t - 2u_1(t) + 2u_2(t),$$

con las condiciones iniciales

$$u_1(0) = -0.4,\ u_2(0) = -0.6.$$

El método de Runge-Kutta de cuarto orden se utilizará para aproximar la solución de este problema usando $h = 0.1$. Las condiciones iniciales dan $w_{1,0} = -0.4$ y $w_{2,0} = -0.6$. Las ecuaciones (5.48) a (5.51) con $j = 0$ dan

$$k_{1,1} = hf_1(t_0, w_{1,0}, w_{2,0}) = hw_{2,0} = -0.06,$$

$$k_{1,2} = hf_2(t_0, w_{1,0}, w_{2,0}) = h[e^{2t_0}\operatorname{sen} t_0 - 2w_{1,0} + 2w_{2,0}] = -0.04,$$

$$k_{2,1} = hf_1\left(t_0 + \frac{h}{2}, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right) = h[w_{2,0} + \frac{1}{2}k_{1,2}] = -0.062,$$

$$k_{2,2} = hf_2\left(t_0 + \frac{h}{2}, w_{1,0} + \frac{1}{2}k_{1,1}, w_{2,0} + \frac{1}{2}k_{1,2}\right)$$

$$= h\left[e^{2(t_0+0.05)}\operatorname{sen}(t_0 + 0.05) - 2\left(w_{1,0} + \frac{1}{2}k_{1,1}\right) + 2\left(w_{2,0} + \frac{1}{2}k_{1,2}\right)\right]$$

$$= -0.03247644757,$$

$$k_{3,1} = h\left[w_{2,0} + \frac{1}{2}k_{2,2}\right] = -0.06162832238,$$

$$k_{3,2} = h\left[e^{2(t_0+0.05)}\operatorname{sen}(t_0 + 0.05) - 2\left(w_{1,0} + \frac{1}{2}k_{2,1}\right) + 2\left(w_{2,0} + \frac{1}{2}k_{2,2}\right)\right]$$

$$= -0.03152409237,$$

$$k_{4,1} = h[w_{2,0} + k_{3,2}] = -0.06315240924,$$

y

$$k_{4,2} = h[e^{2(t_0+0.1)}\operatorname{sen}(t_0 + 0.1) - 2(w_{1,0} + k_{3,1}) + 2(w_{2,0} + k_{3,2})] = 0.02178637298.$$

Por tanto,

$$w_{1,1} = w_{1,0} + \frac{1}{6}(k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1}) = -0.4617333423 \quad \text{y}$$

$$w_{2,1} = w_{2,0} + \frac{1}{6}(k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2}) = -0.6316312421.$$

El valor $w_{1,1}$ aproxima $u_1(0.1) = y(0.1) = 0.2e^{2(0.1)}[\operatorname{sen} 0.1 - 2\cos 0.1]$ y $w_{2,1}$ aproxima $u_2(0.1) = y'(0.1) = 0.2e^{2(0.1)}(4\operatorname{sen} 0.1 - 3\cos 0.1)$.

En la tabla 5.18 se incluye el conjunto de valores $w_{1,j}$ y $w_{2,j}$ para $j = 0, 1, \ldots, 10$ y se comparan con los valores reales de $u_1(t) = 0.2e^{2t}(\operatorname{sen} t - 2\cos t)$ y de $u_2(t) = u_1'(t) = 0.2e^{2t}(4\operatorname{sen} t - 3\cos t)$. ■

Tabla 5.18

t_j	$y(t_j) = u_1(t_j)$	$w_{1,j}$	$y'(t_j) = u_2(t_j)$	$w_{2,j}$	$\lvert y(t_j) - w_{1,j}\rvert$	$\lvert y'(t_j) - w_{2,j}\rvert$
0.0	−0.40000000	−0.00000000	−6.0000000	−0.6000000	0	0
0.1	−0.46173297	−0.46173334	−0.6316304	−0.63163124	3.7×10^{-7}	7.75×10^{-7}
0.2	−0.52555905	−0.52555988	−0.6401478	−0.64014895	8.3×10^{-7}	1.01×10^{-6}
0.3	−0.58860005	−0.58860144	−0.6136630	−0.61366381	1.39×10^{-6}	8.34×10^{-7}
0.4	−0.64661028	−0.64661231	−0.5365821	−0.53658203	2.03×10^{-6}	1.79×10^{-7}
0.5	−0.69356395	−0.69356666	−0.3887395	0.38873810	2.71×10^{-6}	5.96×10^{-7}
0.6	−0.72114849	−0.72115190	−0.1443834	−0.14438087	3.41×10^{-6}	7.75×10^{-7}
0.7	−0.71814890	−0.71815295	0.2289917	0.22899702	4.05×10^{-6}	2.03×10^{-6}
0.8	−0.66970677	−0.66971133	0.7719815	0.77199180	4.56×10^{-6}	5.30×10^{-6}
0.9	−0.55643814	−0.55644290	0.534764	0.15347815	4.76×10^{-6}	9.54×10^{-6}
1.0	−0.35339436	−0.35339886	2.578741	0.25787663	4.50×10^{-6}	1.34×10^{-5}

También podemos utilizar `dsolve` de Maple con ecuaciones de orden superior. Nótese que la n-ésima derivada $y^{(n)}(t)$ se especifica por medio de `(D@@n)(y)(t)`. Para definir la ecuación diferencial del ejemplo 2, usamos

```
>def2:=(D@@2)(y)(t)-2*D(y)(t)+2*y(t)=exp(2*t)*sen(t);
```

y para especificar las condiciones iniciales usamos

```
> init2:=y(0)=-0.4, D(y)(0)=-0.6;
```

La solución se logra aplicando el comando

```
>sol2:=dsolve({def2,init2},y(t));
```

para obtener

$$sol2 := y(t) = -\frac{2}{5}e^{(2t)}\cos(t) + \frac{1}{5}e^{(2t)}\operatorname{sen}(t)$$

Aislamos la solución en forma de función mediante

```
> g:=rhs(sol2);
```

para obtener $y(1.0) = g(1.0)$, introducimos

```
> evalf(subs(t=1.0,g));
```

que da el resultado −.3533943558.

También se dispone del método de Runge-Kutta-Fehlberg para las ecuaciones de orden superior, a través del comando `dsolve` con la opción numérica. Introducimos el comando

```
>g:=dsolve({def2,init2},y(t),numeric):
```

con la respuesta de Maple

$$g := \text{proc}(rkf45_x)\ \ldots\ \text{fin proc}$$

Podemos aproximar $y(1.0)$ aplicando el comando

```
>g(1.0);
```

para obtener

$$[t = 1.0, y(t) = -.353394346807534676, \frac{\partial}{\partial t}y(t) = 2.57874665940482072]$$

En forma semejante podemos extender los demás métodos de un paso a los sistemas. Si extendemos con control de error los métodos como el de Runge-Kutta-Fehlberg, entonces debemos examinar la exactitud de cada componente de la solución numérica $(w_{1j}, w_{2j}, \ldots, w_{mj})$. Si uno de los componentes no ofrece suficiente exactitud, será necesario recalcular la solución numérica completa $(w_{1j}, w_{2j}, \ldots, w_{mj})$.

Los métodos multipasos y los métodos predictores-correctores también pueden extenderse a los sistemas. Una vez más, si se usa el control de error cada componente debe ser exacto. También el método de extrapolación se puede extender a los sistemas, pero la notación se vuelve extremadamente compleja. Si el lector desea profundizar en este tema, le recomendamos consultar [HNW].

Los teoremas de convergencia y las estimaciones de error de los sistemas se asemejan a los que vimos en la sección 5.10 para ecuaciones individuales, salvo que las cotas están dadas a partir de las normas vectoriales, tema que veremos en el capítulo 7. (Una buena obra de consulta en la que se explican estos teoremas es [Gel, pp. 45-72].)

CONJUNTO DE EJERCICIOS 5.9

1. Aplique el método de Runge-Kutta para sistemas y aproxime con él las soluciones de los siguientes sistemas de ecuaciones diferenciales de primer orden. Después, compare los resultados con las soluciones reales.

   a. $u_1' = 3u_1 + 2u_2 - (2t^2 + 1)e^{2t}, \quad 0 \le t \le 1, \quad u_1(0) = 1;$
   $u_2' = 4u_1 + u_2 + (t^2 + 2t - 4)e^{2t}, \quad 0 \le t \le 1, \quad u_2(0) = 1;$
   $h = 0.2;$ soluciones reales $u_1(t) = \frac{1}{3}e^{5t} - \frac{1}{3}e^{-t} + e^{2t}$ y $u_2(t) = \frac{1}{3}e^{5t} + \frac{2}{3}e^{-t} + t^2e^{2t}$.

   b. $u_1' = -4u_1 - 2u_2 + \cos t + 4 \operatorname{sen} t, \quad 0 \le t \le 2, \quad u_1(0) = 0;$
   $u_2' = 3u_1 + u_2 - 3 \operatorname{sen} t, \quad 0 \le t \le 2, \quad u_2(0) = -1;$
   $h = 0.1;$ soluciones reales $u_1(t) = 2e^{-t} - 2e^{-2t} + \operatorname{sen} t$ y $u_2(t) = -3e^{-t} + 2e^{-2t}$.

   c. $u_1' = u_2, \quad 0 \le t \le 2, \quad u_1(0) = 1;$
   $u_2' = -u_1 - 2e^t + 1, \quad 0 \le t \le 2, \quad u_2(0) = 0;$
   $u_3' = -u_1 - e^t + 1, \quad 0 \le t \le 2, \quad u_3(0) = 1;$
   $h = 0.5;$ soluciones reales $u_1(t) = \cos t + \operatorname{sen} t - e^t + 1$, $u_2(t) = -\operatorname{sen} t + \cos t - e^t$ y $u_3(t) = -\operatorname{sen} t + \cos t$.

   d. $u_1' = u_2 - u_3 + t, \quad 0 \le t \le 1, \quad u_1(0) = 1;$
   $u_2' = 3t^2, \quad 0 \le t \le 1, \quad u_2(0) = 1;$
   $u_3' = u_2 + e^{-t}, \quad 0 \le t \le 1, \quad u_3(0) = -1;$
   $h = 0.1;$ soluciones reales $u_1(t) = -0.05t^5 + 0.25t^4 + t + 2 - e^{-t}$, $u_2(t) = t^3 + 1$ y $u_3(t) = 0.25t^4 + t - e^{-t}$.

2. Use el algoritmo de Runge-Kutta para sistemas y aproxime con él las soluciones de las siguientes ecuaciones diferenciales de orden superior. Compare después los resultados con las soluciones reales.

   a. $y'' - 2y' + y = te^t - t, \quad 0 \le t \le 1, \quad y(0) = y'(0) = 0,$ con $h = 0.1$; solución real $y(t) = \frac{1}{6}t^3 e^t - te^t + 2e^t - t - 2$.

b. $t^2y'' - 2ty' + 2y = t^3 \ln t, \quad 1 \leq t \leq 2, \quad y(1) = 1, \quad y'(1) = 0$, con $h = 0.1$; solución real $y(t) = \frac{7}{4}t + \frac{1}{2}t^3 \ln t - \frac{3}{4}t^3$.

c. $y''' + 2y'' - y' - 2y = e^t, \quad 0 \leq t \leq 3, \quad y(0) = 1, \quad y'(0) = 2, \quad y''(0) = 0$, con $h = 0.2$; solución real $y(t) = \frac{43}{36}e^t + \frac{1}{4}e^{-t} - \frac{4}{9}e^{-2t} + \frac{1}{6}te^t$.

d. $t^3y''' - t^2y'' + 3ty' - 4y = 5t^3 \ln t + 9t^3, \quad 1 \leq t \leq 2, \quad y(1) = 0, \quad y'(1) = 1, y''(1) = 3$, con $h = 0.1$; solución real $y(t) = -t^2 + t\cos(\ln t) + t\,\text{sen}(\ln t) + t^3 \ln t$.

3. Cambie el algoritmo predictor de cuarto orden de Adams para obtener soluciones aproximadas a los sistemas de ecuaciones de primer orden.

4. Repita el ejercicio 1 usando el algoritmo desarrollado en el ejercicio 3.

5. Repita el ejercicio 2 usando el algoritmo desarrollado en el ejercicio 3.

6. Suponga que el péndulo descrito en el ejemplo inicial de este capítulo mide 2 pies de largo y que $g = 32.17$ pies/s^2. Con $h = 0.1$ s, compare el ángulo θ obtenido en el caso de los dos siguientes problemas de valor inicial cuando $t = 0$, 1 y 2 s:

a. $\dfrac{d^2\theta}{dt^2} + \dfrac{g}{L}\,\text{sen}\,\theta = 0, \quad \theta(0) = \dfrac{\pi}{6}, \quad \theta'(0) = 0,$

b. $\dfrac{d^2\theta}{dt^2} + \dfrac{g}{L}\theta = 0, \quad \theta(0) = \dfrac{\pi}{6}, \quad \theta'(0) = 0.$

7. El estudio de los modelos matemáticos para predecir la dinámica demográfica de especies antagónicas nació con las obras independientes que, en la primera parte del siglo XX, publicaron A. J. Lotka y V. Volterra. Considere el problema de predecir la población de dos especies, una de las cuales es depredadora y cuya población en el tiempo t es $x_2(t)$ y la otra es la presa, cuya población es $x_1(t)$. Supondremos que la presa dispone siempre de suficiente comida y que su natalidad en cualquier momento es proporcional a la cantidad de presas vivas en ese tiempo; es decir, la natalidad (de la presa) es $k_1x_1(t)$. La mortalidad de la presa depende del número de presas y de depredadores vivos en ese tiempo. Para simplificar los cálculos, supondremos que la mortalidad (de la presa) es $= k_2x_1(t)x_2(t)$. En cambio, la natalidad del depredador depende del suministro de comida $x_1(t)$ y también del número de depredadores que intervienen en el proceso de reproducción. Por tal razón, suponemos que la natalidad (de los depredadores) es $k_3x_1(t)x_2(t)$. Supondremos que su mortalidad es proporcional a la cantidad de depredadores vivos en el tiempo; es decir, mortalidad (de los depredadores) $= k_4x_2(t)$.

Dado que $x_1'(t)$ y $x_2'(t)$ representan, respectivamente, el cambio de las poblaciones de presas y depredadores en el tiempo, el problema se expresa mediante el sistema de ecuaciones diferenciales no lineales.

$$x_1'(t) = k_1x_1(t) - k_2x_1(t)x_2(t) \quad \text{y} \quad x_2'(t) = k_3x_1(t)x_2(t) - k_4x_2(t).$$

Resuelva este sistema para $0 \leq t \leq 4$, suponiendo que la población inicial de la presa es de 1000 y la de los depredadores es de 500, y que las constantes son $k_1 = 3$, $k_2 = 0.002$, $k_3 = 0.0006$ y $k_4 = 0.5$. Dibuje una gráfica de las soluciones de este problema, graficando ambas poblaciones con el tiempo y describa los fenómenos físicos representados. ¿Tiene este modelo demográfico una solución estable? De ser así, ¿con que valores de x_1 y x_2 es estable la solución?

8. En el ejercicio 7 consideramos el problema de predecir la población por medio de un modelo de depredador-presa. Otro problema de este tipo se refiere a dos especies que compiten por la misma comida. Si con $x_1(t)$ y $x_2(t)$ denotamos los números de especies vivas en el tiempo t, a menudo se supone que, aunque la natalidad de cada especie es simplemente proporcional al número de animales vivos en ese tiempo, la mortalidad de cada especie depende de la población de ambas especies. Supondremos que la población de un par determinado de especies se describe mediante las ecuaciones:

$$\frac{dx_1(t)}{dt} = x_1(t)[4 - 0.0003x_1(t) - 0.0004x_2(t)]$$

y

$$\frac{dx_2(t)}{dt} = x_2(t)[2 - 0.0002x_1(t) - 0.0001x_2(t)].$$

Si se sabe que la población inicial de cada especie es de 10 000 encuentre la solución de este sistema cuando $0 \leq t \leq 4$. ¿Tiene este modelo demográfico una solución estable? De ser así, ¿con qué valores de x_1 y x_2 es estable la solución?

5.10 Estabilidad

En este capítulo hemos descrito varios métodos que sirven para aproximar la solución de un problema de valor inicial. Aunque existen muchos otros, seleccionamos los anteriores porque generalmente cumplen con tres criterios:

1. Su desarrollo es tan claro que el estudiante que cursa el primer año de análisis numérico puede entender cómo funcionan y por qué dan buenos resultados.
2. Uno o más de los métodos darán resultados satisfactorios en la mayor parte de los problemas que deben resolver los estudiantes de ciencias e ingeniería.
3. Los métodos más avanzados y complejos tienen como base uno de los métodos que hemos expuesto aquí o bien son una combinación de ellos.

En esta sección explicaremos por qué estos métodos dan resultados satisfactorios y no así algunos métodos semejantes a ellos. Antes de empezar la explicación, es necesario incluir dos definiciones referentes a la convergencia de los métodos de ecuaciones en diferencias de un paso a la solución de esa ecuación, a medida que el tamaño de paso disminuye.

Definición 5.18 Se dice que el método de la ecuación en diferencias de un paso con el error local de truncamiento $\tau_i(h)$ en el i-ésimo paso es **consistente** o compatible con la ecuación diferencial que aproxima, si

$$\lim_{h\to 0} \max_{1\leq i\leq N} |\tau_i(h)| = 0. \qquad \blacksquare$$

Observe que esta definición es *local*, pues para cada uno de los valores $\tau_i(h)$ estamos suponiendo que la aproximación w_{i-1} y la solución exacta $y(t_{i-1})$ son iguales. Un medio más realista de analizar los efectos que se producen al hacer pequeño h, consiste en determinar el efecto *global* del método. Éste es el error máximo del método en el intervalo total de la aproximación, suponiendo que el método dé el resultado exacto en el valor inicial.

Definición 5.19 Se dice que un método de la ecuación en diferencias de un paso es **convergente** respecto a la ecuación diferencial que aproxima, si:

$$\lim_{h\to 0} \max_{1\le i\le N} |w_i - y(t_i)| = 0,$$

donde $y_i = y(t_i)$ denota el valor exacto de la solución de la ecuación diferencial y w_i es la aproximación obtenida a partir del método de diferencias en el i-ésimo paso. ■

Al examinar la desigualdad (5.10) de la sección 5.2 en la fórmula de cota de error del método de Euler, según el teorema. 5.9 puede decirse que

$$\max_{1\le i\le N} |w_i - y(t_i)| \le \frac{Mh}{2L} \, |e^{L(b-a)} - 1|.$$

y, por tanto, el método de Euler es convergente respecto a la ecuación diferencial que cumple las condiciones de este teorema y la razón de convergencia es $O(h)$.

Un método de un paso es consistente precisamente cuando la ecuación en diferencias tiende a la ecuación diferencial cuando el tamaño de paso tiende a cero; es decir, el error local de truncamiento se aproxima a cero cuando el tamaño de paso tiende a cero. La definición de convergencia ofrece una connotación semejante. Un método es convergente precisamente cuando la solución de la ecuación de diferencias tiende a la solución de la ecuación diferencial, conforme el tamaño de paso se acerca a cero.

El otro tipo de cota de error del problema que ocurre cuando se emplean los métodos de diferencias para aproximar las soluciones de las ecuaciones diferenciales, se debe a que no se utilizan resultados exactos. En la práctica, ni las condiciones iniciales ni las operaciones aritméticas que se efectúan después están representadas exactamente, debido al error de redondeo asociado a la aritmética de dígitos finitos. En la sección 5.2 vimos que esta consideración puede ocasionar dificultades, incluso en el método convergente de Euler. Para analizar esta situación, al menos parcialmente, trataremos de determinar cuáles métodos son estables, en el sentido de que los cambios o perturbaciones pequeñas en las condiciones iniciales produzcan cambios igualmente pequeños en las aproximaciones posteriores; es decir, un método **estable** es aquel cuyos resultados se basan *continuamente* en los datos iniciales.

El concepto de estabilidad de la ecuación en diferencias de un paso se parece un poco a la condición de una ecuación diferencial bien planteada, por ello no debe sorprendernos que la condición de Lipschitz aparezca aquí, como sucedió en el teorema correspondiente de las ecuaciones diferenciales, teorema 5.6.

El inciso (i) del siguiente teorema se refiere a la estabilidad de un método de un paso. La demostración de este resultado no es difícil y se incluye en el ejercicio 1. El inciso (ii) del teorema 5.20 se refiere a las condiciones suficientes para que un método consistente sea convergente. El inciso (iii) justifica el comentario hecho en la sección 5.5 sobre el control del error global de un método mediante el control de su error local de truncamiento, e implica que, cuando este error tiene la razón de convergencia $O(h^n)$, el error global presentará la misma razón de convergencia. Las demostraciones de los incisos (ii) y (iii) son más difíciles que la demostración del inciso (i), y pueden encontrarse en el material presentado en [Gel, pp. 57-58].

Teorema 5.20 Supóngase que aproximamos el problema de valor inicial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha,$$

mediante un método de diferencias de un paso en la forma

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + h\phi(t_i, w_i, h).$$

Supóngase, además, que existe un número $h_0 > 0$ y que $\phi(t, w, h)$ es continua y satisface la condición de Lipschitz en la variable w con la constante de Lipschitz L en el conjunto

$$D = \{(t, w, h) \mid a \leq t \leq b, -\infty < w < \infty, 0 \leq h \leq h_0\}.$$

Entonces

(i) El método es estable;
(ii) El método de diferencias es convergente si y sólo si es consistente, el cual equivale a

$$\phi(t, y, 0) = f(t, y), \quad \text{para toda } a \leq t \leq b;$$

(iii) Si existe una función τ y, para cada $i = 1, 2, \ldots, N$, el error local de truncamiento $\tau_i(h)$ satisface $|\tau_i(h)| \leq \tau(h)$ siempre que $0 \leq h \leq h_0$ entonces

$$|y(t_i) - w_i| \leq \frac{\tau(h)}{L} e^{L(t_i - a)}.$$ ■

EJEMPLO 1 Considere el método modificado de Euler dado por

$$w_0 = \alpha,$$

$$w_{i+1} = w_i + \frac{h}{2}\left[f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))\right], \quad \text{para } i = 0, 1, \ldots, N-1.$$

Verificaremos que este método satisfaga la hipótesis del teorema 5.20. En este método

$$\phi(t, w, h) = \frac{1}{2} f(t, w) + \frac{1}{2} f(t + h, w + hf(t, w)).$$

Si f cumple la condición de Lipschitz en $\{(t, w) \mid a \leq t \leq b, -\infty < w < \infty\}$ en la variable w con la constante L, entonces como

$$\phi(t, w, h) - \phi(t, \overline{w}, h) = \frac{1}{2} f(t, w) + \frac{1}{2} f(t + h, w + hf(t, w)) - \frac{1}{2} f(t, \overline{w}) - \frac{1}{2} f(t + h, \overline{w} + hf(t, \overline{w})),$$

la condición de Lipschitz en f nos lleva a

$$\begin{aligned}
|\phi(t, w, h) - \phi(t, \overline{w}, h)| &\leq \frac{1}{2} L|w - \overline{w}| + \frac{1}{2} L\,|w + hf(t, w) - \overline{w} - hf(t, \overline{w})| \\
&\leq L|w - \overline{w}| + \frac{1}{2} L|hf(t, w) - hf(t, \overline{w})| \\
&\leq L|w - \overline{w}| + \frac{1}{2} hL^2|w - \overline{w}| \\
&= \left(L + \frac{1}{2} hL^2\right)|w - \overline{w}|.
\end{aligned}$$

Por tanto, ϕ cumple la condición de Lipschitz en w en el conjunto

$$\{(t, w, h) \mid a \leq t \leq b, -\infty < w < \infty, 0 \leq h \leq h_0\}.$$

para cualquier $h_0 > 0$ con la constante

$$L' = L + \frac{1}{2}h_0L^2.$$

Finalmente, si f es continua en $\{(t, w) \mid a \leq t \leq b, -\infty < w < \infty\}$ entonces ϕ será continua en

$$\{(t, w, h) \mid a \leq t \leq b, -\infty < w < \infty, 0 \leq h \leq h_0\}.$$

por tanto, el teorema 5.20 implica que el método modificado de Euler es estable. Usando $h = 0$ tenemos

$$\phi(t, w, 0) = \frac{1}{2}f(t, w) + \frac{1}{2}f(t + 0, w + 0 \cdot f(t, w)) = f(t, w),$$

así que se cumple la condición de consistencia expresada en el teorema 5.20, inciso (ii). El método es, pues, convergente. Más aún, hemos visto que, en este método, el error local de truncamiento es $O(h^2)$, de manera que la convergencia del método modificado de Euler también tiene la razón $O(h^2)$. ■

En los métodos multipasos, los problemas relacionados con la consistencia, la convergencia y la estabilidad se complican aún más, a causa del número de aproximaciones que requiere cada paso. En los métodos de un paso, la aproximación w_{i+1} depende directamente sólo de la aproximación anterior w_i; en cambio, los métodos multipasos usan, al menos, dos de las aproximaciones precedentes, por su parte, los métodos que habitualmente se emplean requieren más aproximaciones.

El método general multipasos con el que se aproxima la solución de los problemas de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha, \tag{5.53}$$

puede escribirse en la forma

$$\begin{aligned} &w_0 = \alpha, \ w_1 = \alpha_1, \ldots, w_{m-1} = \alpha_{m-1}, \\ &w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m}), \end{aligned} \tag{5.54}$$

para cada $i = m - 1, m, \ldots, N - 1$, donde $a_0, a_1, \ldots, a_{m+1}$ son constantes y, como de costumbre, $h = (b - a)/N$ y $t_i = a + ih$.

El error local de truncamiento en un método multipasos expresado en esta forma es

$$\tau_{i+1}(h) = \frac{y(t_{i+1}) - a_{m-1}y(t_i) - \cdots - a_0y(t_{i+1-m})}{h} - F(t_i, h, y(t_{i+1}), y(t_i), \ldots, y(t_{i+1-m})),$$

para cada $i = m - 1, m, \ldots, N - 1$. Al igual que en los métodos de un paso, el error local de truncamiento mide como la solución $y(t)$ de la ecuación diferencial no satisface la ecuación en diferencias.

En el método de Adams-Bashforth de cuatro pasos, hemos visto que

$$\tau_{i+1}(h) = \frac{251}{720} y^{(5)}(\mu_i)h^4, \quad \text{para algún } \mu_i \in (t_{i-3}, t_{i+1}),$$

mientras que el método de Adams-Moulton de tres pasos tiene

$$\tau_{i+1}(h) = -\frac{19}{720} y^{(5)}(\mu_i)h^4, \quad \text{para algún } \mu_i \in (t_{i-2}, t_{i+1}),$$

naturalmente siempre que $y \in C^5[a, b]$.

En todo el análisis haremos dos suposiciones acerca de la función F:

1. Si $f \equiv 0$ (es decir, si la ecuación diferencial es homogénea), entonces también $F = 0$.
2. F satisface la condición de Lipschitz respecto a $\{w_j\}$ en el sentido de que existe una constante L y para cada par de sucesiones $\{v_j\}_{j=0}^N$ y $\{\tilde{v}_j\}_{j=0}^N$ y para $i = m-1, m, \ldots, N-1$, tenemos

$$|F(t_i, h, v_{i+1}, \ldots, v_{i+1-m}) - F(t_i, h, \tilde{v}_{i+1}, \ldots, \tilde{v}_{i+1-m})| \leq L \sum_{j=0}^{m} |v_{i+1-j} - \tilde{v}_{i+1-j}|. \quad \blacksquare$$

El método explícito de Adams-Bashforth y el implícito de Adams-Moulton cumplen con estas condiciones, siempre que f satisfaga la condición de Lipschitz. (Véase el ejercicio 2.)

En los métodos multipasos el concepto de convergencia es el mismo que el de los métodos de un paso; un método multipasos es **convergente**, si la solución de la ecuación de diferencias se aproxima a la solución de la ecuación diferencial, a medida que el tamaño de paso se acerca a cero. Esto significa que $\lim_{h\to 0} \max_{0\leq i\leq N} |w_i - y(t_i)| = 0$.

Sin embargo, en la consistencia se presenta una situación diferente. Una vez más, queremos que un método multipasos sea consistente a condición de que la ecuación de diferencias aproxime la ecuación diferencial, a medida que el tamaño de paso se acerca a cero; es decir, el error local de truncamiento debe aproximarse a cero en cada paso a medida que el tamaño de éste se aproxima a cero. La condición adicional se presenta debido al número de valores iniciales que requiere el método multipasos. Como únicamente el primer valor inicial $w_0 = \alpha$, suele ser exacto, debemos exigir que los errores de todos los valores iniciales $\{\alpha_i\}$ se aproximen a cero, conforme el tamaño de paso se acerca a cero. Por tanto,

$$\lim_{h\to 0} |\tau_i(h)| = 0, \quad \text{para toda } i = m, m+1, \ldots, N \text{ y} \tag{5.55}$$

$$\lim_{h\to 0} |\alpha_i - y(t_i)| = 0, \quad \text{para toda } i = 1, 2, \ldots, m-1, \tag{5.56}$$

deben ser verdaderos para que un método multipasos de la forma (5.54) sea **consistente**. Nótese que (5.56) implica que un método multipasos no será consistente, salvo que también el método de un paso que genera los valores iniciales lo sea.

El siguiente teorema de métodos multipasos se asemeja al teorema 5.20, inciso (iii), y nos da la relación existente entre el error local de truncamiento y el error global de un método multipasos. Ofrece la justificación teórica para intentar controlar el error global regulando el error local de truncamiento. En [IK, pp. 387-388] se encuentra la demostración de una forma ligeramente más general de este teorema.

Teorema 5.21 Suponga que aproximamos el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

por medio del método explícito predictor-corrector de Adams, con la ecuación predictora de m pasos de Adams-Bashforth

$$w_{i+1} = w_i + h[b_{m-1}f(t_i, w_i) + \cdots + b_0 f(t_{i+1-m}, w_{i+1-m})],$$

con el error local de truncamiento $\tau_{i+1}(h)$ y una ecuación implícita correctora de $(m-1)$ pasos de Adams-Moulton

$$w_{i+1} = w_i + h\left[\tilde{b}_{m-1}f(t_i, w_{i+1}) + \tilde{b}_{m-2}f(t_i, w_i) + \cdots + \tilde{b}_0 f(t_{i+2-m}, w_{i+2-m})\right],$$

con el error local de truncamiento $\tilde{\tau}_{i+1}(h)$. Además, supóngase que $f(t, y)$ y $f_y(t, y)$ son continuas en $D = \{(t, y) \mid a \leq t \leq b \text{ y en } -\infty < y < \infty\}$ y que f_y está acotada. Entonces, el error local de truncamiento $\sigma_{i+1}(h)$ del método predictor-corrector es

$$\sigma_{i+1}(h) = \tilde{\tau}_{i+1}(h) + \tau_{i+1}\tilde{b}_{m-1}\frac{\partial f}{\partial y}(t_{i+1}, \theta_{i+1}),$$

donde θ_{i+1} es un número entre cero y $h\tau_{i+1}(h)$.

Más aún, existen las constantes k_1 y k_2 tales que

$$|w_i - y(t_i)| \leq \left[\max_{0 \leq j \leq m-1} |w_j - y(t_j)| + k_1\sigma(h)\right]e^{k_2(t_i-a)},$$

donde $\sigma(h) = \max_{m \leq j \leq N} |\sigma_j(h)|$. ■

Antes de explicar las relaciones existentes entre consistencia, convergencia y estabilidad en los métodos multipasos, debemos examinar más detenidamente la ecuación de diferencias para un método multipasos. Al hacerlo, descubriremos por qué escogimos los métodos de Adams como nuestros métodos multipasos estándar.

Con la ecuación en diferencias (5.54)

$$w_0 = \alpha, \; w_1 = \alpha_1, \ldots, \; w_{m-1} = \alpha_{m-1},$$
$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0 w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m}),$$

que se incluyó al iniciar esta exposición se relaciona el polinomio característico del método, dado por

$$P(\lambda) = \lambda^m - a_{m-1}\lambda^{m-1} - a_{m-2}\lambda^{m-2} - \ldots - a_1\lambda - a_0. \tag{5.57}$$

Las magnitudes de las raíces de la ecuación característica de un método multipasos se asocian a la estabilidad del método respecto al error de redondeo. Para entender esto, vamos a aplicar el método multipasos estándar (5.54) al problema trivial de valor inicial

$$y' \equiv 0, \quad y(a) = \alpha, \quad \text{donde } \alpha \neq 0. \tag{5.58}$$

Este problema tiene la solución exacta $y(t) \equiv \alpha$. Al examinar las ecuaciones (5.26) y (5.27) en la sección 5.6, observamos que, en teoría, cualquier método multipasos produce la so

lución exacta $w_n = \alpha$ para toda n. La única desviación respecto a la solución exacta se debe al error de redondeo intrínseco asociado a los cálculos que requiere el método.

El lado derecho de la ecuación diferencial en (5.58) tiene $f(t, y) = 0$, de modo que, conforme a la suposición (1), tendremos $F(t_i, h, w_{i+1}, w_{i+2}, \ldots, w_{i+1-m}) = 0$ en la ecuación de diferencias (5.54). En consecuencia, la forma normal de esta última ecuación se convierte en

$$w_{i+1} = a_{m-1} w_i + a_{m-2} w_{i-1} + \cdots + a_0 w_{i+1-m}. \tag{5.59}$$

Supongamos que λ es una de las raíces de la ecuación característica asociadas con (5.54). Entonces $w_n = \lambda^n$ para cada n es una solución de (5.59) porque

$$\lambda^{i+1} - a_{m-1}\lambda^i - a_{m-2}\lambda^{i-1} - \cdots - a_0\lambda^{i+1-m} = \lambda^{i+1-m}[\lambda^m - a_{m-1}\lambda^{m-1} - \cdots - a_0] = 0.$$

De hecho, si $\lambda_1, \lambda_2, \ldots, \lambda_m$ son raíces distintas de la ecuación característica en (5.54), podemos demostrar que toda solución de (5.59) puede expresarse como

$$w_n = \sum_{i=1}^{m} c_i \lambda_i^n, \tag{5.60}$$

para un conjunto único de constantes $c_1, c_2, \ldots, c_m$.

Puesto que la solución exacta de (5.58) es $y(t) = \alpha$ la elección $w_n = \alpha$ para toda n es una solución de (5.59). Al aplicar este hecho en (5.59), obtenemos

$$0 = \alpha - \alpha a_{m-1} - \alpha a_{m-2} - \cdots - \alpha a_0 = \alpha[1 - a_{m-1} - a_{m-2} - \cdots - a_0].$$

Lo anterior significa que $\lambda = 1$ es una de las soluciones de la ecuación característica (5.57). Supondremos que en la representación (5.60) esta solución está descrita por $\lambda_1 = 1$ y $c_1 = \alpha$, así que todas las soluciones de (5.59) se expresan como

$$w_n = \alpha + \sum_{i=2}^{m} c_i \lambda_i^n. \tag{5.61}$$

Si todos los cálculos fueran exactos, las constantes $c_2, c_3, \ldots, c_m$ serían cero. En la práctica, las constantes $c_2, c_3, \ldots, c_m$ no son cero debido al error de redondeo. De hecho, este error crece de manera exponencial a menos que $|\lambda_i| \leq 1$ para las raíces $\lambda_2, \lambda_3, \ldots, \lambda_m$. Cuanto menor sea la magnitud de estas raíces, más estable será el método respecto al crecimiento del error de redondeo.

Al deducir (5.61) hicimos la suposición simplificadora de que todas las raíces de la ecuación característica son distintas. La situación se asemeja a la que se presenta cuando se tienen raíces múltiples. Por ejemplo, si $\lambda_k = \lambda_{k+1} = \cdots = \lambda_{k+p}$ para alguna k y p, esto simplemente requiere reemplazar la suma

$$c_k\lambda_k^n + c_{k+1}\lambda_{k+1}^n + \cdots + c_{k+p}\lambda_{k+p}^n$$

en (5.61) por medio de

$$c_k\lambda_k^n + c_{k+1}n\lambda_k^{n-1} + c_{k+2}n(n-1)\lambda_k^{n-2} + \cdots + c_{k+p}[n(n-1)\cdots(n-p+1)]\lambda_k^{n-p}. \tag{5.62}$$

(Véase [He2, pp. 119-145].) Aunque la forma de la solución está modificada, el efecto del redondeo sigue creciendo exponencialmente si $|\lambda_k| > 1$.

A pesar de haber considerado sólo el caso especial de aproximar los problemas de valor inicial de la forma (5.58), las características de estabilidad de esta ecuación determinan la estabilidad de la situación cuando $f(t, y)$ no es idénticamente cero. Ello se debe al hecho de que la solución de la ecuación homogénea (5.58) está integrada en la solución de cualquier ecuación. Esta explicación da origen a las siguientes definiciones.

Definición 5.22 Denotemos con $\lambda_1, \lambda_2, \ldots, \lambda_m$, las raíces (no necesariamente distintas) de la ecuación característica

$$P(\lambda) = \lambda^m - a_{m-1}\lambda^{m-1} - \cdots - a_1\lambda - a_0 = 0$$

asociadas al método multipasos de diferencias

$$w_0 = \alpha, \quad w_1 = \alpha_1, \ldots, \quad w_{m-1} = \alpha_{m-1}$$

y

$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m}).$$

Si $|\lambda_i| \leq 1$ para cada $i = 1, 2, \ldots, m$ y si todas las raíces con valor absoluto 1 son raíces simples, entonces se dice que el método de diferencia cumple la **condición de raíz**. ■

Definición 5.23

(i) Se da el nombre de métodos **fuertemente estables** a los que cumplen la condición de raíz y tienen $\lambda = 1$ como la única raíz de la ecuación característica de magnitud uno.

(ii) Se da el nombre de métodos **débilmente estables** a los que cumplen la condición de raíz y tienen más de una raíz distinta de magnitud uno.

(iii) Se da el nombre de métodos **inestables** a los que no cumplen la condición de raíz. ■

La consistencia y la convergencia de un método multipasos se relacionan estrechamente con la estabilidad de redondeo del método. En el teorema siguiente se incluyen en forma detallada estas conexiones. Consúltese en [IK, pp. 410-417] la demostración de este resultado y la teoría en que se basa.

Teorema 5.24 Un método multipasos de la forma

$$w_0 = \alpha, \quad w_1 = \alpha_1, \ldots, \quad w_{m-1} = \alpha_{m-1},$$

donde

$$w_{i+1} = a_{m-1}w_i + a_{m-2}w_{i-1} + \cdots + a_0w_{i+1-m} + hF(t_i, h, w_{i+1}, w_i, \ldots, w_{i+1-m})$$

es estable si y sólo si cumple la condición de raíz. Además, si el método de diferencia es consistente con la ecuación diferencial, entonces el método será estable si y sólo si es convergente. ■

EJEMPLO 2 Hemos visto que el método de Adams-Bashforth de cuarto orden puede expresarse como

$$w_{i+1} = w_i + hF(t_i, h, w_{i+1}, \ldots, w_{i-3}),$$

donde:

$$F(t_i, h, w_{i+1}, w_i, \ldots, w_{i-3}) = \frac{h}{24}[55 f(t_i, w_i) - 59 f(t_{i-1}, w_{i-1})$$

$$+ 37 f(t_{i-2}, w_{i-2}) - 9 f(t_{i-3}, w_{i-3})];$$

así que $m = 4$, $a_0 = 0$, $a_1 = 0$, $a_2 = 0$ y $a_3 = 1$.

En consecuencia, la ecuación característica de este método de Adams-Bashforth es

$$0 = P(\lambda) = \lambda^4 - \lambda^3 = \lambda^3 (\lambda - 1)$$

que tiene las raíces $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_3 = 0$ y $\lambda_4 = 0$. Cumple con la condición de raíz y es fuertemente estable.

El método de Adams-Moulton tiene una ecuación característica semejante, $p(\lambda) = \lambda^3 - \lambda^2$, con raíces $\lambda_1 = 1$, $\lambda_2 = 0$ y $\lambda_3 = 0$, y también es fuertemente estable. ■

EJEMPLO 3 En la sección 5.6 se presentó el método explícito multipaso dado por

$$w_{i+1} = w_{i-3} + \frac{4h}{3}[2 f(t_i, w_i) - f(t_{i-1}, w_{i-1}) + 2 f(t_{i-2}, w_{i-2})]$$

como el método explícito de Milne. Puesto que la ecuación característica de este método, $P(\lambda) = \lambda^4 - 1 = 0$, tiene cuatro raíces de magnitud uno $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = i$ y $\lambda_4 = -i$ el método cumple la condición de raíz, pero sólo es débilmente estable.

Consideremos el problema de valor inicial

$$y' = -6y + 6, \quad 0 \le t \le 1, \quad y(0) = 2,$$

que tiene la solución exacta $y(t) = 1 + e^{-6t}$. Con fines de comparación, usamos el método explícito fuertemente estable de Adams-Bashforth de cuarto orden y el método de Milne para aproximar la solución de este problema cuando $h = 0.1$, con los valores exactos para los valores iniciales. Los resultados incluidos en la tabla 5.19 muestran los efectos de un método débilmente estable en comparación con los de un método fuertemente estable en este problema.

Tabla 5.19

t_i	Valores exactos $y(t_i)$	Método de Adams-Bashforth w_i	Error $\lvert y_i - w_i \rvert$	Método de Milne w_i	Error $\lvert y_i - w_i \rvert$
0.10000000		1.5488116		1.5488116	
0.20000000		1.3011942		1.3011942	
0.30000000		1.1652989		1.1652989	
0.40000000	1.0907180	1.0996236	8.906×10^{-3}	1.0983785	7.661×10^{-3}
0.50000000	1.0497871	1.0513350	1.548×10^{-3}	1.0417344	8.053×10^{-3}
0.60000000	1.0273237	1.0425614	1.524×10^{-2}	1.0486438	2.132×10^{-2}
0.70000000	1.0149956	1.0047990	1.020×10^{-2}	0.9634506	5.154×10^{-2}
0.80000000	1.0082297	1.0359090	2.768×10^{-2}	1.1289977	1.208×10^{-1}
0.90000000	1.0045166	0.9657936	3.872×10^{-2}	0.7282684	2.762×10^{-1}
1.00000000	1.0024788	1.0709304	6.845×10^{-2}	1.6450917	6.426×10^{-1}

En la sección 5.6 elegimos el método de Adams-Bashforth-Moulton como nuestro método predictor-corrector estándar de cuarto orden sobre el método de Milne-Simpson

del mismo orden, porque los métodos de Adams-Bashforth y de Adams-Moulton son fuertemente estables. Tienden más a dar aproximaciones exactas con una clase más amplia de problemas que el método predictor-corrector que se basa en los procedimientos de Milne y Simpson, los cuales son débilmente estables. ■

CONJUNTO DE EJERCICIOS 5.10

1. Para demostrar el teorema 5.20, inciso (i), pruebe que las hipótesis implican que existe una constante $K > 0$ tal que

$$|u_i - v_i| \leq K|u_0 - v_0|, \quad \text{para cada } 1 \leq i \leq N,$$

siempre que $\{u_i\}_{i=1}^N$ y que $\{v_i\}_{i=1}^N$ satisfagan la ecuación en diferencias $w_{i+1} = w_i + h\phi(t_i, w_i, h)$.

2. En los métodos de Adams-Bashforth y de Adams-Moulton de cuarto orden,
   a. Demuestre que, si $f = 0$ entonces

$$F(t_i, h, w_{i+1}, \ldots, w_{i+1-m}) = 0.$$

   b. Demuestre que si f cumple la condición de Lipschitz con la constante L, entonces existe una constante C con

$$|F(t_i, h, w_{i+1}, \ldots, w_{i+1-m}) - F(t_i, h, v_{i+1}, \ldots, v_{i+1-m})| \leq C\sum_{j=0}^{m}|w_{i+1-j} - v_{i+1-j}|.$$

3. Use los resultados del ejercicio 17 de la sección 5.4 para demostrar que el método de Runge-Kutta de cuarto orden es consistente.

4. Considere la ecuación diferencial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha.$$

   a. Demuestre que

$$y'(t_i) = \frac{-3y(t_i) + 4y(t_{i+1}) - y(t_{i+2})}{2h} + \frac{h^2}{3}y'''(\xi_i),$$

   para alguna ξ_i donde $t_i < \xi_i < t_{i+2}$.

   b. El inciso (a) sugiere el método de diferencias

$$w_{i+2} = 4w_{i+1} - 3w_i - 2hf(t_i, w_i), \quad \text{para } i = 0, 1, \ldots, N-2.$$

   Use este método para resolver

$$y' = 1 - y, \quad 0 \leq t \leq 1, \quad y(0) = 0,$$

   con $h = 0.1$. Utilice los valores iniciales $w_0 = 0$ y $w_1 = y(t_1) = 1 - e^{-0.1}$.

   c. Repita el inciso (b) con $h = 0.01$ y con $w_1 = 1 - e^{-0.01}$.

   d. Analice la consistencia, estabilidad y convergencia de este método.

5. Dado el método multipasos

$$w_{i+1} = -\frac{3}{2}w_i + 3w_{i-1} - \frac{1}{2}w_{i-2} + 3hf(t_i, w_i), \quad \text{para } i = 2, \ldots, N-1,$$

con los valores iniciales w_0, w_1, w_2:

a. Obtenga el error local de truncamiento.

b. Explique la consistencia, estabilidad y convergencia.

6. Obtenga una solución aproximada de la ecuación diferencial

$$y' = -y, \quad 0 \leq t \leq 10, \quad y(0) = 1,$$

aplicando el método de Milne con $h = 0.1$ y luego con $h = 0.01$ dados los valores iniciales $w_0 = 1$ y $w_1 = e^{-h}$ en ambos casos. ¿Cómo afecta la disminución de h de $h = 0.1$ a $h = 0.01$ al número de dígitos correctos en las soluciones aproximadas en $t = 1$ y $t = 10$?

7. Investigue la estabilidad del método de diferencias

$$w_{i+1} = -4w_i + 5w_{i-1} + 2h[f(t_i, w_i) + 2hf(t_{i-1}, w_{i-1})],$$

para $i = 1, 2, \ldots, N-1$ con los valores iniciales w_0, w_1.

8. Considere el problema $y' = 0$ para $0 \leq t \leq 10$ con $y(0) = 0$ que tiene la solución $y \equiv 0$. Si se aplica al problema el método de diferencia del ejercicio 4, entonces

$$w_{i+1} = 4w_i - 3w_{i-1}, \quad \text{para } i = 1, 2, \ldots, N-1,$$

$$w_0 = 0 \quad \text{y} \quad w_1 = \alpha_1.$$

Suponga que $w_1 = \alpha_1 = \varepsilon$ donde ε es un pequeño error de redondeo. Calcule exactamente w_i para $i = 2, 3, \ldots, 6$ a fin de descubrir cómo se propaga el error ε.

5.11 Ecuaciones diferenciales rígidas

Todos los métodos para aproximar la solución de problemas de valor inicial tienen términos de error que implican una derivada de orden superior de la solución a la ecuación. Si la derivada puede acotarse de manera razonable, entonces el método tendrá una cota predecible para el error, que puede usarse para estimar la precisión de la aproximación. Aun cuando la derivada crezca al aumentar el número de pasos, el error podrá controlarse de manera relativa, siempre que la magnitud de la solución también aumente. Sin embargo, con frecuencia surgen problemas cuando la magnitud de la derivada crece, pero la solución no. En este caso, el error puede crecer tanto que domine los cálculos. Los problemas de valor inicial a los que probablemente les ocurra lo anterior se llaman **ecuaciones rígidas** y son bastante comunes, particularmente en el estudio de vibraciones, reacciones químicas y circuitos eléctricos. Los sistemas rígidos reciben su nombre del movimiento en sistemas de masa-resorte que tienen constantes de resorte grandes.

Las ecuaciones diferenciales rígidas se caracterizan como aquéllas cuya solución exacta tienen un término de la forma e^{-ct}, donde c es una constante positiva grande. Por lo general, esto sólo es parte de la solución, llamada solución *transitoria*. La parte más importante de la solución es la solución de *estado estacionario*. La parte transitoria de una ecuación rígida decaerá rápidamente a cero al aumentar t, pero como la n-ésima derivada

de este término tiene magnitud $c^n e^{-ct}$, la derivada no decae tan rápido. De hecho, como la derivada en el término del error no se evalúa en t, sino en un número entre cero y t, los términos de derivadas pueden crecer cuando t aumenta (de hecho, pueden hacerlo muy rápidamente). Por fortuna, las ecuaciones rígidas se pueden predecir a partir del problema físico del que se deduce la ecuación y, con cuidado, se puede mantener al error bajo control. La manera de hacer esto se analiza en esta sección.

EJEMPLO 1 El sistema de los problemas de valor inicial

$$u_1' = 9u_1 + 24u_2 + 5\cos t - \frac{1}{3}\operatorname{sen} t, \quad u_1(0) = \frac{4}{3}$$

$$u_2' = -24u_1 - 51u_2 - 9\cos t + \frac{1}{3}\operatorname{sen} t, \quad u_2(0) = \frac{2}{3}$$

tiene la solución única

$$u_1(t) = 2e^{-3t} - e^{-39t} + \frac{1}{3}\cos t,$$

$$u_2(t) = -e^{-3t} + 2e^{-39t} - \frac{1}{3}\cos t.$$

El término transitorio e^{-39t} de la solución hace rígido a este sistema. Al aplicar el algoritmo 5.7, o sea el método de Runge-Kutta de cuarto orden para sistemas, se obtienen los resultados de la tabla 5.20. Cuando $h = 0.05$, se logra estabilidad y las aproximaciones son exactas. Sin embargo, el aumento en el tamaño de paso $h = 0.1$ origina los resultados desastrosos que se muestran en la tabla. ■

Tabla 5.20

t	$u_1(t)$	$w_1(t)$ $h = 0.05$	$w_1(t)$ $h = 0.1$	$u_2(t)$	$w_2(t)$ $h = 0.05$	$w_2(t)$ $h = 0.1$
0.1	1.793061	1.712219	−2.645169	−1.032001	−0.8703152	7.844527
0.2	1.423901	1.414070	−18.45158	−0.8746809	−0.8550148	38.87631
0.3	1.131575	1.130523	−87.47221	−0.7249984	−0.7228910	176.4828
0.4	0.9094086	0.9092763	−934.0722	−0.6082141	−0.6079475	789.3540
0.5	0.7387877	9.7387506	−1760.016	−0.5156575	−0.5155810	3520.00
0.6	0.6057094	0.6056833	−7848.550	−0.4404108	−0.4403558	15697.84
0.7	0.4998603	0.4998361	−34989.63	−0.3774038	−0.3773540	69979.87
0.8	0.4136714	0.4136490	−155979.4	−0.3229535	−0.3229078	311959.5
0.9	0.3416143	0.3415939	−695332.0	−0.2744088	−0.2743673	1390664.
1.0	0.2796748	0.2796568	−3099671.	−0.2298877	−0.2298511	6199352.

Aunque la rigidez suele asociarse a sistemas de ecuaciones diferenciales, las características de aproximación de un método numérico que se aplique a un sistema rígido, pueden predecirse mediante el análisis del error producido cuando aplicamos el método a una simple ecuación de prueba,

$$y' = \lambda y, \quad y(0) = \alpha, \quad \text{donde } \lambda < 0. \tag{5.63}$$

La solución de esta ecuación es $y(t) = \alpha e^{\lambda t}$, que contiene el término transitorio $e^{\lambda t}$; el término de estado estacionario es cero; de ahí que sea fácil determinar las características de

aproximación del método. (Una explicación más completa del error de redondeo asociado a los sistemas rígidos requiere el análisis de la ecuación de prueba cuando λ es un número complejo con parte imaginaria negativa; consúltese [Ge1, p. 222].)

Primero consideremos la aplicación del método de Euler a la ecuación de prueba. Con $h = (b - a)/N$ y $t_j = jh$ para $j = 0, 1, 2, \ldots, N$, la ecuación (5.8) implica que

$$w_0 = \alpha,$$

y que

$$w_{j+1} = w_j + h(\lambda w_j) = (1 + h\lambda)w_j,$$

así que

$$w_{j+1} = (1 + h\lambda)^{j+1}\, w_0 = (1 + h\lambda)^{j+1}\alpha, \quad \text{para } j = 0, 1, \ldots, N - 1. \tag{5.64}$$

Puesto que la solución exacta es $y(t) = \alpha e^{\lambda t}$ el error absoluto será

$$|y(t_j) - w_j| = |e^{jh\lambda} - (1 + h\lambda)^j|\,|\alpha| = |(e^{h\lambda})^j - (1 + h\lambda)^j|\,|\alpha|,$$

y la exactitud depende de la precisión con que el término $1 + h\lambda$ aproxime $e^{h\lambda}$. Cuando $\lambda < 0$, la solución exacta $(e^{h\lambda})^j$ decae a cero al crecer j, pero de acuerdo con (5.64) la aproximación tendrá esta propiedad sólo si $|1 + h\lambda| < 1$. En el método de Euler, esto realmente restringe el tamaño de paso h para satisfacer $h < 2/|\lambda|$.

Supongamos ahora que se introduce un error de redondeo δ_0 en la condición inicial para el método de Euler,

$$w_0 = \alpha + \delta_0.$$

En el j-ésimo paso para el error de redondeo es

$$\delta_j = (1 + h\lambda)^j\, \delta_0.$$

Dado que $\lambda < 0$, la condición del control del crecimiento del error de redondeo es igual a la del control del error absoluto $|1 + h\lambda| < 1$ lo que implica $h < 2/|\lambda|$.

La situación es semejante en otros métodos de un paso. En general, existe una función Q con la propiedad de que el método de diferencias, al ser aplicado a la ecuación de prueba, nos da

$$w_{i+1} = Q(h\lambda)w_i. \tag{5.65}$$

La exactitud del método se basa en la precisión con que $Q(h\lambda)$ aproxime $e^{h\lambda}$ y el error crecerá sin cota si $|Q(h\lambda)| > 1$. Un método de Taylor de n-ésimo orden, por ejemplo, tendrá estabilidad respecto al crecimiento del error de redondeo y del error absoluto, a condición de que escojamos h para que satisfaga

$$\left|1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \cdots + \frac{1}{n!}h^n\lambda^n\right| < 1.$$

En el ejercicio 6 se examina el caso específico en que el método es el método clásico de Runge-Kutta de cuarto orden, o sea el método de Taylor de orden cuatro.

Cuando se aplica un método multipasos como el de la forma (5.54) a la ecuación de prueba, el resultado es

$$w_{j+1} = a_{m-1}w_j + \cdots + a_0 w_{j+1-m} + h\lambda(b_m w_{j+1} + b_{m-1}w_j + \cdots + b_0 w_{j+1-m}),$$

para $j = m - 1, \ldots, N - 1$, o bien

$$(1 - h\lambda b_m)w_{j+1} - (a_{m-1} + h\lambda b_{m-1})w_j - \cdots - (a_0 + h\lambda b_0)w_{j+1-m} = 0.$$

A esta ecuación de diferencias homogéneas se asocia un **polinomio característico**

$$Q(z, h\lambda) = (1 - h\lambda b_m)z^m - (a_{m-1} + h\lambda b_{m-1})z^{m-1} - \cdots - (a_0 + h\lambda b_0).$$

Este polinomio es semejante al polinomio característico (5.57), pero también incorpora la ecuación de prueba. En este caso la teoría nos recuerda la explicación sobre la estabilidad dada en la sección 5.10.

Supongamos que se dan $w_0, \ldots, w_{m-1}$ y que, para $h\lambda$ fijo, sean $\beta_1, \ldots, \beta_m$ las raíces de la ecuación $Q(z, h\lambda) = 0$. Si $\beta_1, \ldots, \beta_m$ son distintos, entonces existen las constantes $c_1, \ldots, c_m$ con

$$w_j = \sum_{k=1}^{m} c_k(\beta_k)^j, \quad \text{para } j = 0, \ldots, N. \tag{5.66}$$

Si $Q(z, h\lambda) = 0$ tiene raíces múltiples, w_j se define de manera semejante. [Véase la ecuación (5.62) en la sección 5.10.] Si queremos que w_j aproxime exactamente $y(t_j) = e^{jh\lambda} = (e^{h\lambda})^j$ entonces todas las raíces β_k habrán de satisfacer $|\beta_k| < 1$; de lo contrario, algunas opciones de α producirán $c_k \neq 0$ y el término $c_k(\beta_k)^j$ no decaerá a cero.

EJEMPLO 2 La ecuación diferencial de prueba

$$y' = -30y, \quad 0 \le t \le 1.5, \quad y(0) = \frac{1}{3}$$

tiene la solución exacta $y = \frac{1}{3}e^{-30t}$. Cuando se usa $h = 0.1$ para el algoritmo 5.1 de Euler, el algoritmo de Runge-Kutta de cuarto orden 5.2 y el algoritmo predictor-corrector de Adams 5.4, se obtienen los resultados para $t = 1.5$ que se ven en la tabla 5.21. ■

Tabla 5.21

Solución exacta	9.54173×10^{-21}
Método de Euler	-1.09225×10^{4}
Método de Runge-Kutta	3.95730×10^{1}
Método predictor-corrector	8.03840×10^{5}

Las inexactitudes del ejemplo 2 se deben al hecho de que $|Q(h\lambda)| > 1$ para el método de Euler y el de Runge-Kutta y a que $Q(z, h\lambda)$ tiene raíces con módulo mayor que 1 en el método predictor-corrector. Si se quieren aplicar estos métodos al problema, es necesario reducir el tamaño de paso. La siguiente definición sirve para describir la reducción del tamaño de paso que se requiere.

Definición 5.25 **La región R de la estabilidad absoluta** en un método de un paso es $R = \{h\lambda \in \mathcal{C} \mid |Q(h\lambda)| < 1\}$ y en un método multipasos es $R = \{h\lambda \in \mathcal{C} \mid |\beta_k| < 1$ para todas las raíces β_k de $Q(z, h\lambda)\}$. ■

Las ecuaciones (5.65) y (5.66) implican que sólo es posible aplicar eficientemente un método a una ecuación rígida, si $h\lambda$ se encuentra en la región de estabilidad absoluta de él, lo cual en un problema dado sustituye la restricción impuesta al tamaño de h. Aun cuando en la solución exacta el término exponencial decae rápidamente a cero, λh deberá permanecer dentro de la región de estabilidad absoluta a lo largo del intervalo de t valores, para que la aproximación decaiga a cero y el crecimiento de error se mantenga bajo control. Ello significa que, aunque normalmente podríamos aumentar h debido a consideraciones de error de truncamiento, el criterio de estabilidad absoluta hace que h siga siendo pequeño. Los métodos del tamaño variable de paso son muy vulnerables a este problema, pues un examen del error local de truncamiento podría indicar la posibilidad de aumentar el tamaño de paso, lo cual colocaría a λh inadvertidamente fuera de la región de estabilidad absoluta.

La región de estabilidad absoluta de un método suele ser el factor más importante en la obtención de aproximaciones exactas con los sistemas rígidos; por ello, se buscan métodos numéricos cuya región de estabilidad absoluta sea lo más extensa posible. Se dice que un método numérico es **A-estable** si su región R de estabilidad absoluta contiene a todo el semiplano izquierdo.

El **método implícito del trapecio,** dado por

$$\begin{aligned} w_0 &= \alpha, \\ w_{j+1} &= w_j + \frac{h}{2}\left[f(t_{j+1}, w_{j+1}) + f(t_j, w_j)\right], \quad 0 \le j \le N-1, \end{aligned} \tag{5.67}$$

es un método A-estable (véase el ejercicio 9) y es el único método multipasos A-estable. Aunque el método del trapecio no proporciona aproximaciones exactas con tamaños de paso grandes, los errores no crecen exponencialmente.

Los procedimientos que comúnmente se emplean con los sistemas rígidos son los métodos multipasos implícitos. En general, w_{i+1} se obtiene resolviendo iterativamente una ecuación no lineal o un sistema no lineal, con frecuencia mediante el método de Newton. Consideremos, por ejemplo, el método implícito del trapecio

$$w_{j+1} = w_j + \frac{h}{2}\left[f(t_{j+1}, w_{j+1}) + f(t_j, w_j)\right].$$

Después de calcular t_j, t_{j+1} y w_j debemos determinar $w = w_{j+1}$ es decir, la solución de

$$F(w) = w - w_j - \frac{h}{2}\left[f(t_{j+1}, w) + f(t_j, w_j)\right] = 0. \tag{5.68}$$

Para aproximar esta solución, seleccionamos $w_{j+1}^{(0)}$ generalmente como w_j y generamos $w_{j+1}^{(k)}$ al aplicar el método de Newton a (5.68),

$$\begin{aligned} w_{j+1}^{(k)} &= w_{j+1}^{(k-1)} - \frac{F(w_{j+1}^{(k-1)})}{F'(w_{j+1}^{(k-1)})} \\ &= w_{j+1}^{(k-1)} - \frac{w_{j+1}^{(k-1)} - w_j - \frac{h}{2}\left[f(t_j, w_j) + f(t_{j+1}, w_{j+1}^{(k-1)})\right]}{1 - \frac{h}{2} f_y(t_{j+1}, w_{j+1}^{(k-1)})} \end{aligned}$$

hasta que $\left|w_{j+1}^{(k)} - w_{j+1}^{(k-1)}\right|$ sea suficientemente pequeño. Este procedimiento es el que se usa en el algoritmo 5.8. Normalmente sólo se requieren tres o cuatro iteraciones por paso.

El método de la secante puede usarse como una alternativa al método de Newton en la ecuación (5.68), pero entonces se requerirán dos aproximaciones iniciales distintas a w_{j+1}. Cuando se quiere utilizar el método de la secante, se acostumbra usar $w_{j+1}^{(0)} = w_j$ y obtener $w_{j+1}^{(1)}$ a partir de algún método multipasos explícito. Cuando interviene un sistema de ecuaciones rígidas, se requiere una generalización con el método de Newton o de la secante. En el capítulo 10 estudiaremos estos métodos.

ALGORITMO 5.8

Método del trapecio con iteración de Newton

Para aproximar la solución del problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha$$

en $(N + 1)$ números uniformemente espaciados en el intervalo $[a, b]$:

ENTRADA extremos a, b; entero N; condición inicial α; tolerancia TOL; número máximo de iteraciones M en cualquiera de los pasos.

SALIDA aproximaciones w a y en los $(N + 1)$ valores de t o bien un mensaje de falla.

Paso 1 Tome $h = (b - a)/N$;
$t = a$;
$w = \alpha$;
SALIDA (t, w).

Paso 2 Para $i = 1, 2, \ldots, N$ haga pasos 3-7.

Paso 3 Tome $k_1 = w + \frac{h}{2} f(t, w)$;
$w_0 = k_1$;
$j = 1$;
$BAND = 0$.

Paso 4 Mientras $BAND = 0$ haga pasos 5-6.

Paso 5 tome $w = w_0 - \dfrac{w_0 - \frac{h}{2} f(t + h, w_0) - k_1}{1 - \frac{h}{2} f_y(t + h, w_0)}$.

Paso 6 Si $|w - w_0| < TOL$ entonces tome $BAND = 1$
si no, tome $j = j + 1$;
$w_0 = w$;
Si $j > M$ entonces
SALIDA ('Número máximo de iteraciones rebasado');
PARAR.

Paso 7 Tome $t = a + ih$;
SALIDA (t, w).

Paso 8 PARAR. ■

EJEMPLO 3 El problema rígido de valor inicial

$$y' = 5e^{5t}(y - t)^2 + 1, \quad 0 \le t \le 1, \quad y(0) = -1$$

tiene la solución $y(t) = t - e^{-5t}$. Para demostrar los efectos de la rigidez, se aplican el método implícito del trapecio y el Runge-Kutta de cuarto orden con $N = 4$ y $h = 0.25$ y con $N = 5$ y $h = 0.20$. El método del trapecio da buenos resultados en ambos casos usando $M = 10$ y $TOL = 10^{-6}$ igual que cuando se aplica el de Runge-Kutta con $h = 0.2$. Sin embargo, $h = 0.25$ se encuentra fuera de la región de estabilidad absoluta del método de Runge-Kutta, lo cual se observa claramente en los resultados de la tabla 5.22. ■

Tabla 5.22

	Método de Runge-Kutta		Método trapezoidal	
	$h = 0.2$		$h = 0.2$	
t_i	w_i	$\lvert y(t_i) - w_i \rvert$	w_i	$\lvert y(t_i) - w_i \rvert$
0.0	−1.0000000	0	−1.0000000	0
0.2	−0.1488521	1.9027×10^{-2}	−0.1414969	2.6383×10^{-2}
0.4	0.2684884	3.8237×10^{-3}	0.2748614	1.0197×10^{-2}
0.6	0.5519927	1.7798×10^{-3}	0.5539828	3.7700×10^{-3}
0.8	0.7822857	6.0131×10^{-4}	0.7830720	1.3876×10^{-3}
1.0	0.9934905	2.2845×10^{-4}	0.9937726	5.1050×10^{-4}
	$h = 0.25$		$h = 0.25$	
t_i	w_i	$\lvert y(t_i) - w_i \rvert$	w_i	$\lvert y(t_i) - w_i \rvert$
0.0	−1.0000000	0	−1.0000000	0
0.25	0.4014315	4.37936×10^{-1}	0.0054557	4.1961×10^{-2}
0.5	3.4374753	3.01956×10^{0}	0.4267572	8.8422×10^{-3}
0.75	1.44639×10^{23}	1.44639×10^{23}	0.7291528	2.6706×10^{-3}
1.0	Sobreflujo		0.9940199	7.5790×10^{-4}

Hasta aquí, hemos visto sólo un poco de lo que el lector encontrará con frecuencia y debe saber sobre sistemas de ecuaciones diferenciales rígidas, para un estudio más profundo consulte [Ge2], [Lam] o [SGe].

CONJUNTO DE EJERCICIOS 5.11

1. Resuelva los siguientes problemas rígidos de valor inicial aplicando el método de Euler y después compare los resultados con la solución real.
   a. $y' = -9y, \quad 0 \le t \le 1, \quad y(0) = e$, con $h = 0.1$; solución real $y(t) = e^{1-9t}$.
   b. $y' = -20(y - t^2) + 2t, \quad 0 \le t \le 1, \quad y(0) = \frac{1}{3}$, con $h = 0.1$; solución real $y(t) = t^2 + \frac{1}{3}e^{-20t}$.
   c. $y' = -20y + 20 \operatorname{sen} t + \cos t, \quad 0 \le t \le 2, \quad y(0) = 1$, con $h = 0.25$; solución real $y(t) = \operatorname{sen} t + e^{-20t}$.

d. $y' = \frac{50}{y} - 50y, \quad 0 \le t \le 1, \quad y(0) = \sqrt{2}$, con $h = 0.1$; solución real $y(t) = (1 + e^{-100t})^{1/2}$.

2. Repita el ejercicio 1 aplicando el método de Runge-Kutta de cuarto orden.
3. Repita el ejercicio 1 aplicando el método predictor-corrector de cuarto orden de Adams.
4. Repita el ejercicio 1 usando el algoritmo del trapecio. Utilice $TOL = 10^{-5}$.
5. Resuelva el siguiente problema rígido de valor inicial aplicando el método de Runge-Kutta con (a) $h = 0.1$ y (b) $h = 0.025$.

$$u_1' = 32u_1 + 66u_2 + \frac{2}{3}t + \frac{2}{3}, \quad 0 \le t \le 0.5, \quad u_1(0) = \frac{1}{3};$$

$$u_2' = -66u_1 - 133u_2 - \frac{1}{3}t - \frac{1}{3}, \quad 0 \le t \le 0.5, \quad u_2(0) = \frac{1}{3}.$$

Compare los resultados con la solución real,

$$u_1(t) = \frac{2}{3}t + \frac{2}{3}e^{-t} - \frac{1}{3}e^{-100t} \quad y \quad u_2(t) = -\frac{1}{3}t - \frac{1}{3}e^{-t} + \frac{2}{3}e^{-100t}.$$

6. Demuestre que el método de Runge-Kutta de cuarto orden,

$$\begin{aligned} k_1 &= hf(t_i, w_i), \\ k_2 &= hf(t_i + h/2, w_i + k_1/2), \\ k_3 &= hf(t_i + h/2, w_i + k_2/2), \\ k_4 &= hf(t_i + h, w_i + k_3), \\ w_{i+1} &= w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \end{aligned}$$

cuando se aplica a la ecuación diferencial $y' = \lambda y$ puede expresarse como

$$w_{i+1} = \left(1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4\right)w_i.$$

7. Explique la consistencia, estabilidad y convergencia en el método implícito del trapecio

$$w_{i+1} = w_i + \frac{h}{2}[f(t_{i+1}, w_{i+1}) + f(t_i, w_i)], \quad \text{para } i = 0, 1, \ldots, N-1,$$

con $w_0 = \alpha$ aplicado a la ecuación diferencial

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha.$$

8. El método regresivo de un paso de Euler se define por medio de

$$w_{i+1} = w_i + hf(t_{i+1}, w_{i+1}), \quad \text{para } i = 0, \ldots, N-1.$$

   a. Demuestre que $Q(h\lambda) = 1/(1 - h\lambda)$ en el método regresivo de Euler.
   b. Aplique el método regresivo de Euler a las ecuaciones diferenciales del ejercicio 1. Use el método de Newton para w_{i+1}.

9. a. Demuestre que el método implícito del trapecio (5.67) es A-estable.
   b. Demuestre que el método regresivo de Euler descrito en el ejercicio 8 es A-estable.

5.12 Reseña de métodos y de software

En este capítulo estudiamos los métodos con que se aproximan las soluciones de los problemas de valor inicial en las ecuaciones diferenciales ordinarias. Iniciamos exponiendo el método más elemental del análisis numérico: el método de Euler. Este procedimiento no fue lo bastante preciso para utilizarse en las aplicaciones, pero es un ejemplo del comportamiento general de otros métodos más poderosos sin preocuparse por el momento por los problemas algebraicos. Después estudiamos los métodos de Taylor como generalizaciones del de Euler. Descubrimos que son exactos pero complicados, pues hay que determinar extensas derivadas parciales de la ecuación diferencial. Las fórmulas de Runge-Kutta simplifican los métodos de Taylor, pero sin aumentar considerablemente el error. Hasta ahora habíamos considerado exclusivamente los métodos de un paso, los cuales sólo usan datos en el punto calculado más reciente.

En la sección 5.6 describimos los métodos multipasos, junto con los de tipo explícito de Adams-Bashforth y los de tipo implícito de Adams-Moulton. A través de ellos llegamos a los métodos predictores-correctores que usan un método implícito, como el de Adams-Moulton, para corregir la aproximación.

En la sección 5.9 dimos ejemplos de estos procedimientos, que pueden servir para resolver problemas de valor inicial de orden superior y sistemas de problemas de valor inicial.

Los métodos adaptativos más precisos se basan en las técnicas relativamente sencillas de uno y varios pasos. En particular, en la sección 5.5 vimos que el método de Runge-Kutta-Fehlberg es un procedimiento de un paso que trata de seleccionar los espaciamientos de red para mantener bajo control la aproximación. El método predictor-corrector que se presentó en la sección 5.7 se basa en el de Adams-Bashforth de cuatro pasos y en el de Adams-Moulton de tres pasos. También cambia el tamaño de paso para que el error local no rebase determinada tolerancia. El método de extrapolación que se explicó en la sección 5.8 se basa en una modificación del método de punto medio e incorpora la extrapolación para conservar la exactitud deseada de la aproximación.

El último tema del capítulo se refiere a la dificultad que conlleva la aproximación de la solución de una ecuación rígida, o sea una ecuación diferencial cuya solución exacta contiene una parte de la forma $e^{-\lambda t}$, donde λ es una constante positiva. Debe tenerse mucho cuidado con este tipo de problemas, pues de lo contrario los resultados pueden incluir el error de redondeo.

Los métodos del tipo de Runge-Kutta-Fehlberg suelen ser suficientes en los problemas no rígidos en que se requiere una precisión moderada. En los problemas no rígidos en los que se necesita mucha precisión, se recomiendan los métodos de extrapolación. Finalmente, en los problemas no rígidos de valor inicial se emplean extensiones del método implícito del trapecio a los métodos de orden variable y a los métodos implícitos de tamaño variable de paso de tipo Adams.

La biblioteca IMSL incluye dos subrutinas que sirven para aproximar las soluciones de los problemas de valor inicial. Cada método resuelve un sistema de m ecuaciones de primer orden en m variables. Las ecuaciones tienen la forma

$$\frac{du_i}{dt} = f_i(t, u_1, u_2, \ldots, u_m), \quad \text{para } i = 1, 2, \ldots, m.$$

donde $u_i(t_0)$ está dado para cada i. La subrutina de tamaño variable de paso IVPRK se basa en los métodos de Runge-Kutta-Verner de quinto y sexto orden, que se describen en el ejercicio 4 de la sección 5.5. Una subrutina de tipo Adams que se emplea con las ecuaciones rígidas fue ideada por C. William Gear, y está dada por IVPAG. Esta subrutina utiliza

los métodos multipasos implícitos hasta de orden 12 y las fórmulas de diferenciación regresiva hasta de orden 5.

Los procedimientos de tipo Runge-Kutta contenidos en la biblioteca NAG reciben el nombre de D02BGF, D02BHF, D02PCF y D02PDF, D02BGF y D02BHF se basan en la versión Merson del método de Runge-Kutta. En el procedimiento D02CJF está contenido un método de orden variable y de tamaño variable de paso de Adams. En el procedimiento D02EJF está contenido un método de orden variable y de diferencia regresiva de tamaño variable de paso para sistemas rígidos. Otras rutinas incorporan los mismos métodos, pero los iteran hasta que una componente de la solución alcance determinado valor o hasta que una función de la solución sea cero.

La biblioteca netlib incluye varias subrutinas para aproximar las soluciones de problemas de valor inicial en el paquete ODE, localizado en http://www.netlib.org/ode. La subrutina dverk.f se basa en los métodos de quinto y sexto orden de Runge-Kutta-Verner. La subrutina rkf45.f se basa en los métodos de cuarto y quinto orden de Runge-Kutta-Fehlberg descritos en la sección 5.5. Para los problemas de valor inicial con ecuaciones diferenciales ordinarias rígidas, se puede usar la subrutina epsode.f, basada en la fórmula de derivación regresiva con coeficientes variables.

Hay muchos libros que se especializan en la resolución numérica de los problemas de valor inicial. Dos obras clásicas son la de Henrici [He1] y la de Gear [Ge1]. Otros libros que ofrecen un panorama de este campo de estudio son los de Botha y Pinder [BP], Ortega y Poole [OP], Golub y Ortega [GO], Shampine [Sh] y Dormand [Do]. Dos libros de Hairer, Nörsett y Warner ofrecen explicaciones exhaustivas sobre los problemas no rígidos [HNW1] y rígidos [HNW2]. En el libro de Burrage [Bur] se describen los métodos paralelos y secuenciales.

CAPÍTULO

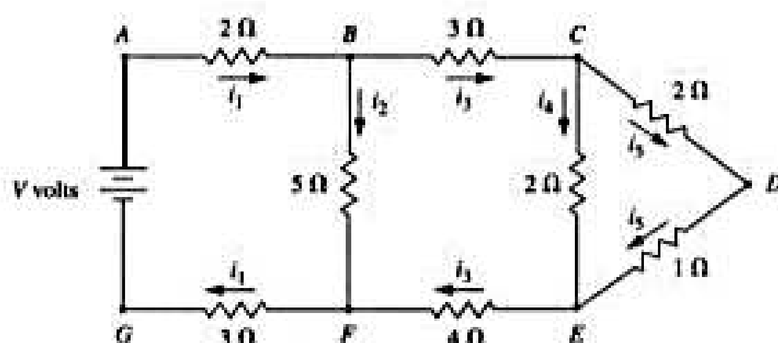
6

Métodos directos para resolver sistemas lineales

▪ ▪ ▪

Las leyes de Kirchhoff de los circuitos eléctricos establecen que el flujo neto de la corriente que pasa por las uniones de un circuito es cero, y que la caída de voltaje neto alrededor de las partes cerradas del circuito también es cero. Supóngase que aplicamos un potencial de V volts entre los puntos A y G en el circuito situado debajo, y que i_1, i_2, i_3, i_4 e i_5 representan el flujo de corriente como se muestra en el diagrama. Al utilizar G como punto de referencia, las leyes de Kirchhoff establecen que las corrientes satisfacen el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} 5i_1 + 5i_2 &= V, \\ i_3 - i_4 - i_5 &= 0, \\ 2i_4 - 3i_5 &= 0, \\ 5i_2 - 7i_3 - 2i_4 &= 0. \end{aligned}$$



En este capítulo estudiaremos la solución a este tipo de sistema. Esta aplicación se explica en el ejercicio 23 de la sección 6.6.

Los sistemas de ecuaciones lineales se utilizan en muchos problemas de ingeniería y de las ciencias, así como en aplicaciones de las matemáticas a las ciencias sociales y al estudio cuantitativo de problemas de administración y economía.

En este capítulo se examinan métodos directos con que se resuelve el sistema lineal

$$
\begin{aligned}
E_1:&\quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1,\\
E_2:&\quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2,\\
&\quad\vdots\\
E_n:&\quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n,
\end{aligned}
\tag{6.1}
$$

Para $x_1, \ldots, x_n$, dadas las a_{ij} con $i, j = 1, 2, \ldots, n$ y b_i, para $i = 1, 2, \ldots, n$. Estas técnicas directas son métodos que proporcionan una respuesta en un número fijo de pasos y sólo están sujetos a los errores de redondeo. En el desarrollo del tema, también presentaremos algunos conceptos básicos tomados del álgebra lineal.

En el capítulo 7 trataremos de los métodos con que se aproxima la solución de los sistemas lineales por medio de métodos iterativos.

6.1 Sistemas de ecuaciones lineales

Utilizamos tres operaciones para simplificar el sistema lineal que se incluye en (6.1):

1. La ecuación E_i puede multiplicarse por una constante λ distinta de cero y la ecuación resultante se emplea en vez de E_i. Esta operación se denota por $(\lambda E_i) \rightarrow (E_i)$.
2. La ecuación E_j puede multiplicarse por cualquier constante λ y sumarse a la ecuación E_i, la ecuación resultante se emplea en vez de E_i. Esta operación se denota por $(E_i + \lambda E_j) \rightarrow (E_i)$.
3. El orden de las ecuaciones E_i y E_j puede intercambiarse. Esta operación se denota por $(E_i) \leftrightarrow (E_j)$.

Con la serie de operaciones que acabamos de incluir, podemos transformar un sistema lineal en otro que puede resolverse más fácilmente con las mismas soluciones. Ilustramos esto en el siguiente ejemplo.

EJEMPLO 1 Las cuatro ecuaciones

$$
\begin{aligned}
E_1:&\quad x_1 + x_2 \phantom{{}-x_3} + 3x_4 = 4,\\
E_2:&\quad 2x_1 + x_2 - x_3 + x_4 = 1,\\
E_3:&\quad 3x_1 - x_2 - x_3 + 2x_4 = -3,\\
E_4:&\quad -x_1 + 2x_2 + 3x_3 - x_4 = 4,
\end{aligned}
\tag{6.2}
$$

se resolverán para x_1, x_2, x_3 y x_4. Primero utilizamos la ecuación E_1 para eliminar la incógnita x_1 en E_2, E_3 y E_4, al efectuar $(E_2 - 2E_1) \rightarrow (E_2)$, $(E_3 - 3E_1) \rightarrow (E_3)$ y $(E_4 + E_1) \rightarrow (E_4)$. El sistema resultante es

$$\begin{aligned}
E_1:&\quad x_1 + x_2 - 3x_4 = 4,\\
E_2:&\quad - x_2 - x_3 - 5x_4 = -7,\\
E_3:&\quad - 4x_2 - x_3 - 7x_4 = -15,\\
E_4:&\quad 3x_2 + 3x_3 + 2x_4 = 8,
\end{aligned}$$

donde, por razones de simplicidad, volvemos a marcar las nuevas ecuaciones con E_1, E_2, E_3 y E_4.

En el sistema nuevo, usamos E_2 para eliminar x_2 en E_3 y E_4 al efectuar $(E_3 - 4E_2) \rightarrow (E_3)$ y $(E_4 + 3E_2) \rightarrow (E_4)$, lo cual da por resultado el sistema

$$\begin{aligned}
E_1:&\quad x_1 + x_2 + 3x_4 = 4,\\
E_2:&\quad - x_2 - x_3 - 5x_4 = -7,\\
E_3:&\quad 3x_3 - 13x_4 = 13,\\
E_4:&\quad - 13x_4 = -13.
\end{aligned} \tag{6.3}$$

El sistema de ecuaciones (6.3) presenta ahora una **forma triangular** (o **reducida**) y puede resolverse para las incógnitas mediante un **proceso de sustitución hacia atrás**. Nótese que E_4 implica que $x_4 = 1$, E_3 puede resolverse para x_3 y dar

$$x_3 = \frac{1}{3}(13 - 13x_4) = \frac{1}{3}(13 - 13) = 0.$$

Y continuando el proceso E_2 nos da

$$x_2 = -(-7 + 5x_4 + x_3) = -(-7 + 5 + 0) = 2,$$

y E_1 da

$$x_1 = 4 - 3x_4 - x_2 = 4 - 3 - 2 = -1.$$

Por tanto, la solución de (6.3) y de (6.2) es $x_1 = -1$, $x_2 = 2$, $x_3 = 0$ y $x_4 = 1$. ■

Al realizar los cálculos del ejemplo 1, no fue necesario escribir todas las ecuaciones completas en cada paso ni retener las variables x_1, x_2, x_3 y x_4 en los cálculos, pues siempre permanecieron en la misma columna. La única variante de un sistema a otro se presentó en los coeficientes de las incógnitas y en los valores del lado derecho de las ecuaciones. Por tal razón, a menudo un sistema lineal se reemplaza con una *matriz*, que contiene toda la información sobre el sistema necesaria para determinar su solución, aunque en una forma compacta.

Definición 6.1 Una **matriz** $(n \times m)$ es un arreglo rectangular de elementos con n renglones y m columnas, donde no sólo es importante el valor de un elemento, sino también su posición en el arreglo. ■

La notación de una matriz de $n \times m$ será una letra mayúscula, como A, para designar la matriz y letras minúsculas con subíndices dobles, como a_{ij}, para indicar la entrada en la intersección del i-ésimo renglón y la j-ésima columna, es decir,

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

EJEMPLO 2 La matriz

$$A = \begin{bmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{bmatrix}$$

es una matriz 2×3 con $a_{11} = 2$, $a_{12} = -1$, $a_{13} = 7$, $a_{21} = 3$, $a_{22} = 1$ y $a_{23} = 0$. ■

La matriz $1 \times n$

$$A = [a_{11}\; a_{12}\; \ldots\; a_{1n}]$$

Recibe el nombre de **vector renglón *n*-dimensional** y una matriz de $n \times 1$

$$A = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}$$

recibe el nombre de **vector columna *n*-dimensional.** Por lo regular, en los vectores se omiten los subíndices innecesarios y en la notación se emplean letras minúsculas negritas. Por tanto,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

denota un vector columna, y

$$\mathbf{y} = [y_1 \quad y_2 \quad \ldots \quad y_n]$$

denota un vector renglón.

Una matriz de $n \times (n + 1)$ puede utilizarse para representar el sistema lineal

$$\begin{aligned} a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n &= b_1, \\ a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n &= b_2, \\ &\vdots \\ a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n &= b_n, \end{aligned}$$

construyendo primero

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

y luego combinando estas matrices para formar la **matriz aumentada**

$$[A, \mathbf{b}] = \left[\begin{array}{cccc:c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right],$$

donde con la línea punteada vertical se separan los coeficientes de las incógnitas de los valores situados en el lado derecho de las ecuaciones.

Al repetir las operaciones descritas en el ejemplo 1 usando la notación matricial, se considera primera la matriz aumentada:

$$\left[\begin{array}{rrrr:r} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right].$$

Si realizamos las operaciones descritas en ese ejemplo se obtienen las matrices

$$\left[\begin{array}{rrrr:r} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right] \quad \text{y} \quad \left[\begin{array}{rrrr:r} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right]$$

Ahora podemos transformar la última matriz en su correspondiente sistema lineal y obtener así las soluciones para x_1, x_2, x_3 y x_4. El método que incluye este proceso se denomina **eliminación gaussiana con sustitución hacia atrás.**

El procedimiento general de eliminación gaussiana que se aplica al sistema lineal

$$\begin{aligned} E_1&: \quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ E_2&: \quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ &\qquad\vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ E_n&: \quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n, \end{aligned} \tag{6.4}$$

se efectúa en forma parecida. Primero formamos la matriz aumentada $\tilde{A}$:

$$\tilde{A} = [A, \mathbf{b}] = \left[\begin{array}{cccc:c} a_{11} & a_{12} & \cdots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & \cdots & a_{2n} & a_{2,n+1} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & a_{n,n+1} \end{array}\right], \tag{6.5}$$

donde A denota la matriz formada por los coeficientes. Los elementos en la $(n + 1)$-ésima columna son los valores de **b**; es decir, $a_{i,n+1} = b_i$ para toda $i = 1, 2, \ldots, n$.

Siempre que $a_{11} \neq 0$, las operaciones correspondientes a $(E_j - (a_{j1}/a_{11})E_1) \to (E_j)$ se efectúan por cada $j = 2, 3, \ldots, n$ para eliminar el coeficiente de x_1 en cada uno de estos renglones. Aunque se espera que los elementos de los renglones $2, 3, \ldots, n$ cambien para facilitar la notación denotamos nuevamente con a_{ij} el elemento del i-ésimo renglón y la j-ésima columna. Teniendo presente esto, aplicamos un procedimiento secuencial cuando $i = 2, 3, \ldots, n - 1$ y realizamos la operación $(E_j - (a_{ji}/a_{ii})E_1) \to (E_j)$ para toda $j = i + 1, i + 2, \ldots, n$, a condición de que $a_{ii} \neq 0$. Con ello se suprime (es decir, se transforma en cero el coeficiente) x_i en cada renglón debajo del i-ésimo para todos los valores de $i = 1, 2, \ldots, n - 1$. La matriz resultante tiene la forma

$$\tilde{\tilde{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & \vdots & a_{1,n+1} \\ 0 & a_{22} & \cdots & a_{2n} & \vdots & a_{2,n+1} \\ \vdots & \ddots & \ddots & & \vdots & \vdots \\ 0 & \cdots & 0 & a_{nn} & \vdots & a_{n,n+1} \end{bmatrix},$$

donde, salvo en el primer renglón no se espera que los valores de a_{ij} concuerden con los de la matriz original $\tilde{A}$ La matriz $\tilde{\tilde{A}}$ representa un sistema lineal con el mismo conjunto solución que el sistema original (6.4). Dado que el nuevo sistema lineal es triangular

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= a_{1,n+1}, \\ a_{22}x_2 + \cdots + a_{2n}x_n &= a_{2,n+1}, \\ \ddots \qquad \vdots \qquad & \quad \vdots \\ a_{nn}x_n &= a_{n,n+1}, \end{aligned}$$

podemos efectuar la sustitución hacia atrás. Al resolver la n-ésima ecuación para x_n obtenemos

$$x_n = \frac{a_{n,n+1}}{a_{nn}}.$$

Al resolver la $(n - 1)$-ésima ecuación para x_{n-1} y al utilizar la incógnita x_n obtenemos

$$x_{n-1} = \frac{a_{n-1,n+1} - a_{n-1,n}x_n}{a_{n-1,n-1}},$$

y, al continuar este proceso, obtenemos

$$x_i = \frac{a_{i,n+1} - a_{i,n}x_n - a_{i,n-1}x_{n-1} - \cdots - a_{i,i+1}x_{i+1}}{a_{ii}} = \frac{a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}},$$

para cada $i = n - 1, n - 2, \ldots, 2, 1$.

El procedimiento de eliminación gaussiana puede ser descrito con mayor precisión, aunque esto supone más complejidad, formando una sucesión de matrices aumentadas $\tilde{A}^{(1)}, \tilde{A}^{(2)}, \ldots, \tilde{A}^{(n)}$, donde $\tilde{A}^{(1)}$ es la matriz $\tilde{A}$ incluida en (6.5) y $\tilde{A}^{(k)}$ para cada $k = 2, 3, \ldots, n$ tiene los elementos $a_{ij}^{(k)}$, donde:

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)}, & \text{cuando } i = 1, 2, \ldots, k - 1 \text{ y } j = 1, 2, \ldots, n + 1, \\ 0, & \text{cuando } i = k, k + 1, \ldots, n \text{ y } j = 1, 2, \ldots, k - 1, \\ a_{ij}^{(k-1)} - \dfrac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)}, & \text{cuando } i = k, k + 1, \ldots, n \text{ y } j = k, k + 1, \ldots, n + 1. \end{cases}$$

Por tanto,

$$\tilde{A}^{(k)} = \left[\begin{array}{ccccccccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} & \vdots & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} & \vdots & a_{2,n+1}^{(2)} \\ \vdots & & \ddots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \vdots & & & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} & \vdots & a_{k-1,n+1}^{(k-1)} \\ \vdots & & & & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & \vdots & a_{k,n+1}^{(k)} \\ \vdots & & & & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & \vdots & a_{n,n+1}^{(k)} \end{array}\right] \qquad (6.6)$$

representa el sistema lineal equivalente para el cual la variable x_{k-1} acaba de eliminarse en las ecuaciones $E_k, E_{k+1}, \ldots, E_n$.

El procedimiento fallará si uno de los elementos $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, \ldots, a_{n-1,n-1}^{(n-1)}, a_{nn}^{(n)}$ es cero porque, en este caso, el paso

$$\left(E_i - \frac{a_{i,k}^{(k)}}{a_{kk}^{(k)}} E_k\right) \rightarrow E_i$$

o no puede efectuarse (lo cual ocurre si uno de $a_{11}^{(1)}, \ldots, a_{n-1,n-1}^{(n-1)}$ es cero) o si no es posible realizar la sustitución hacia atrás (en este caso $a_{nn}^{(n)} = 0$). Lo anterior no significa necesariamente que el sistema no tenga solución, sino que debe modificarse el método para obtenerla. Esto se explica en el siguiente ejemplo.

EJEMPLO 3 Considere el sistema lineal

$$\begin{aligned} E_1&: & x_1 - x_2 + 2x_3 - x_4 &= -8, \\ E_2&: & 2x_1 - 2x_2 + 3x_3 - 3x_4 &= -20, \\ E_3&: & x_1 + x_2 + x_3 &= -2, \\ E_4&: & x_1 - x_2 + 4x_3 + 3x_4 &= 4. \end{aligned}$$

La matriz aumentada es

$$\tilde{A} = \tilde{A}^{(1)} = \left[\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 2 & -2 & 3 & -3 & -20 \\ 1 & 1 & 1 & 0 & -2 \\ 1 & -1 & 4 & 3 & 4 \end{array}\right],$$

y al efectuar las operaciones

$$(E_2 - 2E_1) \rightarrow (E_2), (E_3 - E_1) \rightarrow (E_3) \quad \text{y} \quad (E_4 - E_1) \rightarrow (E_4),$$

obtenemos

$$\tilde{A}^{(2)} = \left[\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

Dado que $a_{22}^{(2)}$, denominado **elemento pivote**, es cero, no podemos continuar el procedimiento en su forma actual. Pero la operación $(E_i) \leftrightarrow (E_j)$ es permitida, por lo cual buscamos los elementos $a_{32}^{(2)}$ y $a_{42}^{(2)}$ en el primer elemento no cero. Puesto que $a_{32}^{(2)} \neq 0$, efectuamos la operación $(E_2) \leftrightarrow (E_3)$ para obtener una nueva matriz.

$$\tilde{A}^{(2)'} = \left[\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 2 & 4 & 12 \end{array}\right].$$

Como ya se eliminó x_2 de E_3 y de E_4, $\tilde{A}^{(3)}$ será $\tilde{A}^{(2)'}$ y continuaremos los cálculos con la operación $(E_4 + 2E_3) \leftrightarrow (E_4)$, que nos da

$$\tilde{A}^{(4)'} = \left[\begin{array}{cccc:c} 1 & -1 & 2 & -1 & -8 \\ 0 & 2 & -1 & 1 & 6 \\ 0 & 0 & -1 & -1 & -4 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right].$$

Finalmente, aplicamos la sustitución hacia atrás:

$$x_4 = \frac{4}{2} = 2,$$

$$x_3 = \frac{[-4 - (-1)x_4]}{-1} = 2,$$

$$x_2 = \frac{[6 - x_4 - (-1)x_3]}{2} = 3,$$

$$x_1 = \frac{[-8 - (-1)x_4 - 2x_3 - (-1)x_2]}{1} = -7.$$ ■

En el ejemplo 2 se explica lo que se hace si $a_{kk}^{(k)} = 0$ para alguna $k = 1, 2, \ldots, n - 1$. La k-ésima columna de $\tilde{A}^{(k-1)}$ proveniente del k-ésimo renglón se analiza en busca del primer elemento no cero. Si $a_{pk}^{(k)} \neq 0$ para alguna p, con $k + 1 \leq p \leq n$, entonces efectuamos la operación $(E_k) \leftrightarrow (E_p)$ para obtener $\tilde{A}^{(k-1)'}$. Después, continuamos el procedimiento para formar $\tilde{A}^{(k)}$, y así sucesivamente. Si $a_{pk}^{(k)} = 0$ para cada p, podemos demostrar (véase el teorema 6.16 en la sección 6.4) que el sistema lineal no tiene solución única y el procedimiento se interrumpe. Finalmente, si $a_{nn}^{(n)} = 0$, el sistema lineal no tiene solución única y el procedimiento vuelve a interrumpirse. En el algoritmo 6.1 se resume la eliminación gaussiana con sustitución hacia atrás. En él se incorpora el pivoteo cuando uno de los pivotes $a_{kk}^{(k)}$ es cero, intercambiando el k-ésimo renglón con el p-ésimo renglón, donde p es el entero más pequeño con valor más grande que k en donde $a_{pk}^{(k)}$ no es cero.

ALGORITMO 6.1

Eliminación gaussiana con sustitución hacia atrás

Para resolver el sistema lineal de $n \times n$

$$
\begin{array}{lccccc}
E_1: & a_{11}x_1 & + a_{12}x_2 & + \cdots & + a_{1n}x_n & = a_{1,n+1} \\
E_2: & a_{21}x_1 & + a_{22}x_2 & + \cdots & + a_{2n}x_n & = a_{2,n+1} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
E_n: & a_{n1}x_1 & + a_{n2}x_2 & + \cdots & + a_{nn}x_n & = a_{n,n+1}
\end{array}
$$

ENTRADA número de incógnitas y ecuaciones n; matriz aumentada $A = (a_{ij})$ donde $1 \le i \le n$ y $1 \le j \le n + 1$.

SALIDA solución $x_1, x_2, \ldots, x_n$ o mensaje de que el sistema lineal no tiene solución única.

Paso 1 Para $i = 1, \ldots, n - 1$ haga pasos 2-4. (*Proceso de eliminación.*)

Paso 2 Sea p el entero más pequeño con $i \le p \le n$ y $a_{pi} \ne 0$.
Si no puede encontrarse un entero p
entonces SALIDA ('no existe solución única');
PARAR.

Paso 3 Si $p \ne i$ entonces realice $(E_p) \leftrightarrow (E_i)$.

Paso 4 Para $j = i + 1, \ldots, n$ haga pasos 5 y 6.

Paso 5 Tome $m_{ji} = a_{ji}/a_{ii}$.

Paso 6 Realice $(E_j - m_{ji} E_i) \to (E_j)$;

Paso 7 Si $a_{nn} = 0$ entonces SALIDA ('no existe solución única')
PARAR.

Paso 8 Tome $x_n = a_{n,n+1}/a_{nn}$. (*Comience la sustitución hacia atrás.*)

Paso 9 Para $i = n - 1, \ldots, 1$ tome $x_i = \left[a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j\right] / a_{ii}$.

Paso 10 SALIDA $(k_1, \ldots, x_n)$; (*Procedimiento terminado exitosamente.*)
PARAR. ■

Todos los sistemas algebraicos por computadora tienen rutinas matriciales. Para definir las matrices y realizar la eliminación gaussiana por medio de Maple, hay que acceder a la biblioteca de álgebra lineal usando el comando.

```
>with(linalg);
```

Para definir la matriz $\tilde{A}^{(1)}$ del ejemplo 2, que designaremos AA, usamos el comando

```
>AA:=matriz(4,5,[1,-1,2,-1,-8,2,-2,3,-3,-20,1,1,1,0,-2,1,-1,4,3,4]);
```

Los dos parámetros, 4 y 5, dan la cantidad de renglones y columnas, respectivamente, y el último parámetro es una lista de elementos de $\tilde{A}^{(1)} = AA$. La función `adrow(AA,i,j,m)`

realiza la operación $(E_j + mE_i) \to (E_j)$ y la función `swaprow(AA,i,j)` realiza la operación $(E_i) \leftrightarrow (E_j)$. Por tanto, la serie de operaciones

```
>AA:=addrow(AA,1,2,-2);
>AA:=addrow(AA,1,3,-1);
>AA:=addrow(AA,1,4,-1);
>AA:=swaprow(AA,2,3);
>AA:=addrow(AA,3,4,2);
```

da la reducción a $\tilde{A}^{(4)}$ que también llamaremos AA. En cambio el comando individual `AA:=gausselim(AA);` devuelve la matriz reducida. La última operación

```
>x:=backsub(AA);
```

da la solución $x := [-7, 3, 2, 2]$.

EJEMPLO 4 Este ejemplo tiene por objeto mostrar lo que sucede cuando el algoritmo 6.1 falla. Los cálculos se harán simultáneamente en dos sistemas lineales:

$$\begin{aligned} x_1 + x_2 + x_3 &= 4, \\ 2x_1 + 2x_2 + x_3 &= 6, \\ x_1 + x_2 + 2x_3 &= 6, \end{aligned} \qquad \text{y} \qquad \begin{aligned} x_1 + x_2 + x_3 &= 4, \\ 2x_1 + 2x_2 + x_3 &= 4, \\ x_1 + x_2 + 2x_3 &= 6. \end{aligned}$$

Estos sistemas dan origen a las matrices

$$\tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 2 & 2 & 1 & 6 \\ 1 & 1 & 2 & 6 \end{array}\right] \text{ y } \tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 2 & 2 & 1 & 4 \\ 1 & 1 & 2 & 6 \end{array}\right]$$

Dado que $a_{11} = 1$, efectuamos $(E_2 - 2E_1) \to (E_2)$ y $(E_3 - E_1) \to (E_3)$ para obtener

$$\tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 1 & 2 \end{array}\right] \text{ y } \tilde{A} = \left[\begin{array}{ccc:c} 1 & 1 & 1 & 4 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

En este punto, $a_{22} = a_{32} = 0$. El algoritmo requiere que el procedimiento se interrumpa, sin que se obtenga la solución a ninguno de los dos sistemas. Al escribir las ecuaciones para cada sistema, obtenemos

$$\begin{aligned} x_1 + x_2 + x_3 &= 4, \\ -x_3 &= -2, \\ x_3 &= 2, \end{aligned} \qquad \text{y} \qquad \begin{aligned} x_1 + x_2 + x_3 &= 4, \\ -x_3 &= -4, \\ x_3 &= 2. \end{aligned}$$

El primer sistema lineal tiene un número infinito de soluciones $x_3 = 2$, $x_2 = 2 - x_1$, con x_1 arbitraria. El segundo sistema da origen a una contradicción $x_3 = 2$, $x_3 = 4$, de manera que no existe solución alguna. Sin embargo, en ambos casos tampoco existe una solución *única*, como concluimos a partir del algoritmo 6.1. ■

Aunque podemos considerar al algoritmo 6.1 como la construcción de las matrices aumentadas $\tilde{A}^{(1)}, \ldots, \tilde{A}^{(n)}$, podemos efectuar los cálculos en una computadora empleando exclusivamente un arreglo de $n \times (n + 1)$ para almacenamiento. En cada paso, simplemente reemplazamos el valor anterior de a_{ij} con el nuevo. Además, podemos guardar los

multiplicadores m_{ji} en los lugares de a_{ji} porque a_{ji} tiene el valor cero para cada $i = 1, 2, \ldots, n-1$, y para cada $j = i+1, i+2, \ldots, n$. Por consiguiente, podemos rescribir A mediante multiplicadores debajo de la diagonal principal y mediante elementos no cero de $\tilde{A}^{(n)}$ arriba de esa diagonal. Según veremos en la sección 6.5, estos valores sirven para resolver otros sistemas lineales que contienen la matriz original A.

Tanto el tiempo necesario para terminar los cálculos como el subsecuente error de redondeo dependen de la cantidad de operaciones de aritmética de punto flotante, que deban efectuarse para resolver un problema de rutina. En términos generales, el tiempo que tardamos en realizar una multiplicación o división en una computadora es más o menos igual, y resulta mucho más largo, que el que tardamos en efectuar una suma o resta. No obstante, las diferencias reales en el tiempo de ejecución dependen del sistema de cómputo que usemos. Para demostrar las operaciones de conteo de un método cualquiera, contaremos las que se requieren para resolver un sistema lineal común de n ecuaciones con n incógnitas empleando el algoritmo 6.1. Debido al diferencial del tiempo, tendremos separado el conteo de las sumas/restas y de las multiplicaciones/divisiones.

En el algoritmo, no se efectúan operaciones aritméticas antes de los pasos 5 y 6. El paso 5 requiere efectuar $(n - i)$ divisiones. La sustitución de la ecuación E_j por $(E_j - m_{ji}E_i)$ en el paso 6 requiere multiplicar m_{ji} por cada término de E_i, lo cual nos da un total de $(n-i)(n-i+1)$ multiplicaciones. Una vez hecho esto, los términos de la ecuación resultante se restan al término correspondiente de E_j. Para ello hay que efectuar $(n-i)(n-i+1)$ restas. Para cada $i = 1, 2, \ldots, n-1$, las operaciones necesarias en los pasos 5 y 6 son las siguientes.

Multiplicaciones/divisiones

$$(n-i) + (n-i)(n-i+1) = (n-i)(n-i+2).$$

Sumas/restas

$$(n-i)(n-i+1).$$

El número total de operaciones que requieren estos pasos se obtiene sumando los conteos de operaciones para cada i. Si recordamos que

$$\sum_{j=1}^{m} 1 = m, \quad \sum_{j=1}^{m} j = \frac{m(m+1)}{2}, \quad \text{y} \quad \sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6},$$

tenemos los siguientes conteos de operaciones.

Multiplicaciones/divisiones

$$\sum_{i=1}^{n-1} (n-i)(n-i+2) = \sum_{i=1}^{n-1} (n^2 - 2ni + i^2 + 2n - 2i)$$

$$= (n^2 + 2n)\sum_{i=1}^{n-1} 1 - 2(n+1)\sum_{i=1}^{n-1} i + \sum_{i=1}^{n-1} i^2 = \frac{2n^3 + 3n^2 - 5n}{6}.$$

Sumas/restas

$$\sum_{i=1}^{n-1}(n-i)(n-i+1)=\sum_{i=1}^{n-1}(n^2-2ni+i^2+n-i)$$

$$=(n^2+n)\sum_{i=1}^{n-1}1-(2n+1)\sum_{i=1}^{n-1}i+\sum_{i=1}^{n-1}i^2=\frac{n^3-n}{3}.$$

Los otros pasos del algoritmo 6.1 que incluyen operaciones aritméticas, son los que se requieren en la sustitución hacia atrás: los pasos 8 y 9. El paso 8 requiere una división. El paso 9 requiere $(n-i)$ multiplicaciones y $(n-i-1)$ sumas en cada término de la suma y después una resta y una división. La cantidad total de operaciones de los pasos 8 y 9 es la siguiente.

Multiplicaciones/divisiones

$$1+\sum_{i=1}^{n-1}((n-i)+1)=\frac{n^2+n}{2}.$$

Sumas/restas

$$\sum_{i=1}^{n-1}((n-i-1)+1)=\frac{n^2-n}{2}.$$

Así pues, el número total de operaciones aritméticas en el algoritmo 6.1 es:

Multiplicaciones/divisiones

$$\frac{2n^3+3n^2-5n}{6}+\frac{n^2+n}{2}=\frac{n^3}{3}+n^2-\frac{n}{3}.$$

Sumas/restas

$$\frac{n^3-n}{3}+\frac{n^2-n}{2}=\frac{n^3}{3}+\frac{n^2}{2}-\frac{5n}{6}.$$

Con n grande, el número total de multiplicaciones y de divisiones es aproximadamente $n^3/3$, igual que el número total de sumas y restas. Así, la cantidad de cálculos y el tiempo necesarios aumentan con n en proporción a n^3, según se observa en la tabla 6.1.

Tabla 6.1

n	Multiplicaciones/divisiones	Sumas/restas
3	17	11
10	430	375
50	44 150	42 875
100	343 300	338 250

CONJUNTO DE EJERCICIOS 6.1

1. En los sistemas lineales siguientes, obtenga, de ser posible, una solución mediante los métodos gráficos. Explique los resultados desde el punto de vista geométrico.

a. $x_1 + 2x_2 = 3,$
$x_1 - x_2 = 0.$

b. $x_1 + 2x_2 = 0,$
$x_1 - x_2 = 0.$

c. $x_1 + 2x_2 = 3,$
$2x_1 + 4x_2 = 6.$

d. $x_1 + 2x_2 = 3,$
$-2x_1 - 4x_2 = 6.$

e. $x_1 + 2x_2 = 0,$
$2x_1 + 4x_2 = 0.$

f. $2x_1 + x_2 = -1,$
$x_1 + x_2 = 2,$
$x_1 - 3x_2 = 5.$

g. $2x_1 + x_2 = -1,$
$4x_1 + 2x_2 = -2,$
$x_1 - 3x_2 = 5.$

h. $2x_1 + x_2 + x_3 = 1,$
$2x_1 + 4x_2 - x_3 = -1.$

2. Use la eliminación gaussiana con sustitución hacia atrás y la aritmética de redondeo a dos dígitos para resolver los sistemas lineales siguientes. No reordene las ecuaciones (la solución exacta de cada sistema es $x_1 = 1$, $x_2 = -1$, $x_3 = 3$.)

a. $4x_1 - x_2 + x_3 = 8,$
$2x_1 + 5x_2 + 2x_3 = 3,$
$x_1 + 2x_2 + 4x_3 = 11.$

b. $4x_1 + x_2 + 2x_3 = 9,$
$2x_1 + 4x_2 - x_3 = -5,$
$x_1 + x_2 - 3x_3 = -9.$

3. Use el algoritmo de la eliminación guassiana para resolver los sistemas lineales siguientes, de ser posible, y determine si se requieren intercambios de renglón:

a. $x_1 - x_2 + 3x_3 = 2,$
$3x_1 - 3x_2 + x_3 = -1,$
$x_1 + x_2 = 3.$

b. $2x_1 - 1.5x_2 + 3x_3 = 1,$
$-x_1 + 2x_3 = 3,$
$4x_1 - 4.5x_2 + 5x_3 = 1.$

c. $2x_1 = 3,$
$x_1 + 1.5x_2 = 4.5,$
$- 3x_2 + 0.5x_3 = -6.6,$
$2x_1 - 2x_2 + x_3 + x_4 = 0.8.$

d. $x_1 - \frac{1}{2}x_2 + x_3 = 4,$
$2x_1 - x_2 - x_3 + x_4 = 5,$
$x_1 + x_2 = 2,$
$x_1 - \frac{1}{2}x_2 + x_3 + x_4 = 5.$

e. $x_1 + x_2 + x_4 = 2,$
$2x_1 + x_2 - x_3 + x_4 = 1,$
$4x_1 - x_2 - 2x_3 + 2x_4 = 0,$
$3x_1 - x_2 - x_3 + 2x_4 = -3.$

f. $x_1 + x_2 + x_4 = 2,$
$2x_1 + x_2 - x_3 + x_4 = 1,$
$-x_1 + 2x_2 + 3x_3 - x_4 = 4,$
$3x_1 - x_2 - x_3 + 2x_4 = -3.$

4. Use el algoritmo de eliminación gaussiana y la aritmética de precisión simple de una computadora para resolver los sistemas lineales siguientes.

a. $\frac{1}{4}x_1 + \frac{1}{5}x_2 + \frac{1}{6}x_3 = 9,$
$\frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = 8,$
$\frac{1}{2}x_1 + x_2 + 2x_3 = 8.$

b. $3.333x_1 + 15920x_2 - 10.333x_3 = 15913,$
$2.222x_1 + 16.71x_2 + 9.612x_3 = 28.544,$
$1.5611x_1 + 5.1791x_2 + 1.6852x_3 = 8.4254.$

c. $x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 + \frac{1}{4}x_4 = \frac{1}{6},$
$\frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 + \frac{1}{5}x_4 = \frac{1}{7},$
$\frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 + \frac{1}{6}x_4 = \frac{1}{8},$
$\frac{1}{4}x_1 + \frac{1}{5}x_2 + \frac{1}{6}x_3 + \frac{1}{7}x_4 = \frac{1}{9}.$

d. $2x_1 + x_2 - x_3 + x_4 - 3x_5 = 7,$
$x_1 + 2x_3 - x_4 + x_5 = 2,$
$-2x_2 - x_3 + x_4 - x_5 = -5,$
$3x_1 + x_2 - 4x_3 + 5x_5 = 6,$
$x_1 - x_2 - x_3 - x_4 + x_5 = 3.$

5. Dado el sistema lineal

$$\begin{aligned} 2x_1 - 6\alpha x_2 &= 3, \\ 3\alpha x_1 - x_2 &= \tfrac{3}{2}. \end{aligned}$$

a. Obtenga el valor o los valores de α para los cuales el sistema no tiene solución.

b. Obtenga el valor o los valores de α para los cuales el sistema tiene una cantidad infinita de soluciones.

c. Suponiendo que existe una solución única para α encuentre la solución.

6. Dado el sistema lineal

$$\begin{aligned} x_1 - x_2 + \alpha x_3 &= -2, \\ -x_1 + 2x_2 - \alpha x_3 &= 3, \\ \alpha x_1 + x_2 + x_3 &= 2. \end{aligned}$$

a. Obtenga el valor o los valores de α para los cuales el sistema no tiene solución.

b. Obtenga el valor o los valores de α para los cuales el sistema tiene una cantidad infinita de soluciones.

c. Suponiendo que hay una solución única para determinada α, obtenga la solución.

7. Demuestre que las operaciones

a. $(\lambda E_i) \to (E_i)$ **b.** $(E_i + \lambda E_j) \to (E_i)$ **c.** $(E_i) \leftrightarrow (E_j)$

no cambian el conjunto de soluciones de un sistema lineal.

8. **Método de Gauss-Jordan**. Este método puede describirse como se indica a continuación. Se usa la i-ésima ecuación para eliminar no sólo x_i en las ecuaciones $E_{i+1}, E_{i+2}, \ldots, E_n$, como se hizo con el método de eliminación gaussiana, sino también en $E_1, E_2, \ldots, E_{i-1}$. Luego de reducir $[A, \mathbf{b}]$ a:

$$\left[\begin{array}{cccc:c} a_{11}^{(1)} & 0 & \ldots & 0 & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots & a_{2,n+1}^{(2)} \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \ldots & 0 & a_{nn}^{(n)} & a_{n,n+1}^{(n)} \end{array}\right],$$

la solución se obtiene usando

$$x_i = \frac{a_{i,n+1}^{(i)}}{a_{ii}^{(i)}},$$

para cada $i = 1, 2, \ldots, n$. Este procedimiento prescinde de la sustitución hacia atrás en la eliminación gaussiana. Construya un algoritmo para el método de Gauss-Jordan configurado a partir del algoritmo 6.1.

9. Use el método de Gauss-Jordan y la aritmética de redondeo a dos dígitos para resolver los sistemas del ejercicio 2.
10. Repita el ejercicio 4 aplicando el método de Gauss-Jordan.
11. a. Demuestre que el método de Gauss-Jordan requiere

$$\frac{n^3}{2} + n^2 - \frac{n}{2} \text{ multiplicaciones/divisiones}$$

y

$$\frac{n^3}{2} - \frac{n}{2} \text{ sumas/restas.}$$

b. Construya una tabla para comparar las operaciones que se requieren con el método de Gauss-Jordan y con el de la eliminación gaussiana para $n = 3, 10, 50, 100$. ¿En cuál de los dos hubo que hacer menos cálculos?

12. Considere el siguiente método híbrido de Gauss-Jordan y de eliminación gaussiana con el cual se resuelve el sistema (6.4). Primero, aplique el método de Gauss-Jordan para reducir el sistema a una forma triangular. Después use la n-ésima ecuación para suprimir los coeficientes de x_n en los $n - 1$ primeros renglones. Una vez hecho esto, utilice la $(n - 1)$-ésima ecuación para suprimir los coeficientes de x_{n-1} en los primeros $n - 2$ renglones, etc. El sistema aparecerá finalmente como el sistema reducido del ejercicio 8.

a. Demuestre que este método requiere

$$\frac{n^3}{3} + \frac{3}{2}n^2 - \frac{5}{6}n \text{ multiplicaciones/divisiones}$$

y

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5}{6}n \text{ sumas/restas.}$$

b. Construya una tabla que compare las operaciones que se requieren con el método de la eliminación gaussiana, el método de Gauss-Jordan y con los métodos híbridos para $n = 3, 10, 50, 100$.

13. Aplique el método híbrido descrito en el ejercicio 12 y la aritmética de redondeo a dos dígitos para resolver los sistemas del ejercicio 2.
14. Repita el ejercicio 4 aplicando el método que se describe en el ejercicio 12.
15. Suponga que en un sistema biológico hay n especies de animales y m fuentes de alimento. Sea x_j la representación de la población de la j-ésima especie para cada $j = 1, \ldots, n$; y sea b_i el suministro disponible del i-ésimo alimento, y con a_{ij} represente la cantidad de la i-ésima comida consumida en promedio por un miembro de la j-ésima especie. El sistema lineal

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1,\\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2,\\
\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \quad &\quad \vdots\\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m,
\end{aligned}$$

representa un equilibrio donde hay suministro diario de comida que satisface exactamente el consumo promedio diario de cada especie.

a. Sea

$$A = (a_{ij}) = \begin{bmatrix} 1 & 2 & 0 & 3 \\ 1 & 0 & 2 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

$\mathbf{x} = (x_j) = [1000, 500, 350, 400]$ y $\mathbf{b} = (b_i) = [3500, 2700, 900]$. ¿Hay suficientes alimentos para satisfacer el consumo promedio diario?

b. ¿Cuál es el número máximo de animales de cada especie que podría agregarse individualmente al sistema de modo que el suministro satisfaciera todavía el consumo.

c. Si se extingue la especie 1, ¿qué aumento individual de cada especie podría soportarse?

d. Si se extingue la especie 2, ¿qué aumento individual de las especies restantes podría soportarse?

16. Una ecuación integral de segunda clase de Fredholm tiene la forma

$$u(x) = f(x) + \int_a^b K(x, t)u(t)dt,$$

donde se dan a y b, así como las funciones f y K. Para aproximar la función u en el intervalo $[a, b]$, se escoge una partición $x_0 = a < x_1 < \cdots x_{m-1} < x_m = b$ y las ecuaciones

$$u(x_i) = f(x_i) + \int_a^b K(x_i, t)u(t)dt, \quad \text{para cada} \quad i = 0, \ldots, m,$$

se resuelven en términos de $u(x_0), u(x_1), \ldots, u(x_m)$. Las integrales se aproximan mediante las fórmulas de cuadraturas que se basan en los nodos $x_0, \ldots, x_m$. En nuestro problema $a = 0$, $b = 1$, $f(x) = x^2$ y $K(x, t) = e^{|x-t|}$.

a. Demuestre que el sistema lineal

$$u(0) = f(0) + \frac{1}{2}[K(0, 0)u(0) + K(0, 1)u(1)],$$

$$u(1) = f(1) + \frac{1}{2}[K(1, 0)u(0) + K(1, 1)u(1)]$$

debe resolverse cuando se aplica la regla del trapecio.

b. Formule y resuelva el sistema lineal que resulta cuando se aplica la regla del trapecio con $n = 4$.

a. Repita el inciso (b) aplicando la regla compuesta de Simpson.

6.2 Estrategias de pivoteo

Al deducir el algoritmo 6.1 descubrimos que cuando uno de los elementos del pivote $a_{kk}^{(k)}$ es cero, se requiere un intercambio de renglones. Este intercambio de renglones presenta la forma $(E_k) \leftrightarrow (E_p)$, donde p es el menor entero mayor que k con $a_{pk}^{(k)} \neq 0$. Si se quiere reducir el error de redondeo, a menudo hay que realizar intercambios de renglones aun cuando los elementos del pivote no sean cero.

Si $a_{kk}^{(k)}$ es de magnitud pequeña en comparación con $a_{jk}^{(k)}$, el multiplicador

$$m_{jk} = \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}}$$

tendrá una magnitud mucho mayor que 1. Los errores de redondeo introducidos en el cálculo de uno de los términos $a_{kl}^{(k)}$ se multiplicarán por m_{jk} cuando se calcule $a_{jl}^{(k+1)}$, lo cual puede incrementar el error inicial. Asimismo, cuando se hace la sustitución hacia atrás con

$$x_k = \frac{a_{k,n+1}^{(k)} - \sum_{j=k+1}^{n} a_{kj}^{(k)}}{a_{kk}^{(k)}},$$

con un valor pequeño de $a_{kk}^{(k)}$, cualquier error del numerador puede aumentar extraordinariamente por la división entre $a_{kk}^{(k)}$. En el siguiente ejemplo se ilustra este problema.

EJEMPLO 1 El sistema lineal

$$\begin{aligned} E_1:&\quad 0.003000x_1 + 59.14x_2 = 59.17 \\ E_2:&\quad 5.291x_1 - 6.130x_2 = 46.78, \end{aligned}$$

tiene la solución exacta $x_1 = 10.00$ y $x_2 = 1.000$. En este sistema realizaremos la eliminación gaussiana mediante la aritmética de redondeo a cuatro dígitos.

El primer elemento del pivote es un número pequeño $a_{11}^{(1)} = 0.003000$ y su multiplicador asociado

$$m_{21} = \frac{5.291}{0.003000} = 1763.6\bar{6},$$

se redondea al número grande 1764. Al realizar $(E_2 - m_{21}E_1) \to (E_2)$ y el redondeo adecuado, obtenemos

$$\begin{aligned} 0.003000x_1 + 59.14x_2 &\approx 59.17 \\ -104300x_2 &\approx 104400, \end{aligned}$$

en vez de los valores precisos,

$$\begin{aligned} 0.003000x_1 + 59.14x_2 &= 59.17 \\ -104309.37\bar{6}x_2 &= -104309.37\bar{6}. \end{aligned}$$

La disparidad de las magnitudes de $m_{21}a_{13}$ y a_{23} ha ocasionado un error de redondeo, pero éste todavía no se ha propagado. La sustitución hacia atrás produce

$$x_2 \approx 1.001,$$

que es una aproximación cercana al valor real, $x_2 = 1.000$. Pero debido al pivote pequeño $a_{11} = 0.003000$,

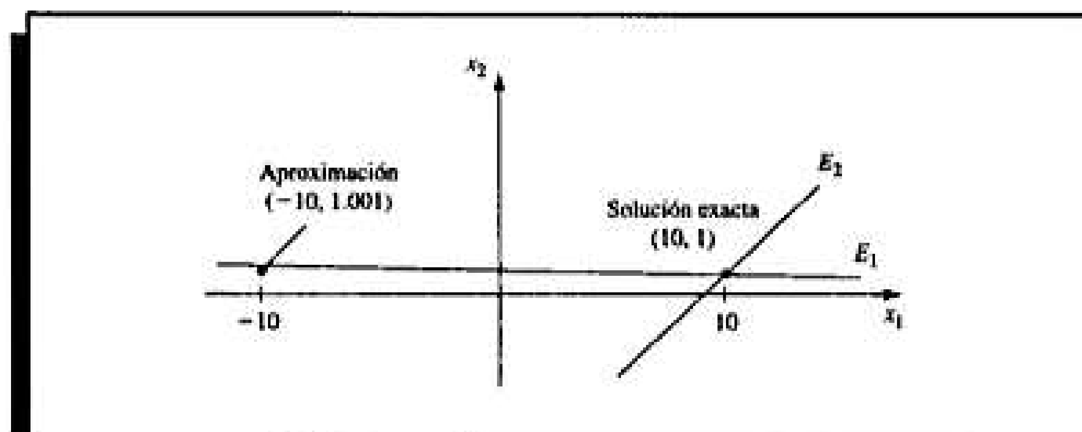
$$x_1 \approx \frac{59.17 - (59.14)(1.001)}{0.003000} = -10.00$$

contiene el pequeño error de 0.001 multiplicado por

$$\frac{59.14}{0.003000} \approx 20\,000.$$

Lo anterior arruina la aproximación al valor real $x_1 = 10.00$ (véase Fig. 6.1). ■

FIGURA 6.1



En el ejemplo 1 observamos los problemas que pueden surgir cuando el elemento pivote $a_{kk}^{(k)}$ es pequeño en comparación con los elementos $a_{ij}^{(k)}$ para $k \leq i \leq n$ y $k \leq j \leq n$. Para evitar este problema empleamos el pivoteo seleccionando un elemento mayor $a_{pq}^{(k)}$ como pivote e intercambiando los renglones k-ésimo y p-ésimo y, en caso necesario, intercambiando después las columnas k-ésima y q-ésima. La estrategia más sencilla consiste en escoger el elemento en la misma columna que está debajo de la diagonal y que tiene el máximo valor absoluto; es decir, determinamos la más pequeña $p \geq k$ tal que

$$|a_{pq}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

y efectuamos $(E_k) \leftrightarrow (E_p)$. En este caso no hay intercambio de columnas.

EJEMPLO 2 Reconsideremos el sistema

$$\begin{aligned} E_1: &\quad 0.003000x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - 6.130x_2 = 46.78. \end{aligned}$$

El procedimiento de pivoteo que acabamos de describir sirve primero para obtener

$$\max\left\{|a_{11}^{(1)}|, |a_{21}^{(1)}|\right\} = \max\{|0.003000|, |5.291|\} = |5.291| = |a_{21}^{(1)}|.$$

Efectuamos la operación $(E_2) \leftrightarrow (E_1)$ para obtener el sistema

$$\begin{aligned} E_1: &\quad 5.291x_1 - 6.130x_2 = 46.78, \\ E_2: &\quad 0.003000x_1 + 59.14x_2 = 59.17. \end{aligned}$$

El multiplicador para este sistema es

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = 0.0005670,$$

y la operación $(E_2 - m_{21}E_1) \to (E_2)$ reduce el sistema

$$\begin{aligned} 5.291x_1 - 6.130x_2 &\approx 46.78, \\ 59.14x_2 &\approx 59.14. \end{aligned}$$

Las respuestas de cuatro dígitos que resultan de la sustitución hacia atrás son los valores correctos $x_1 = 10.00$ y $x_2 = 1.000$. ■

A esta técnica se le llama **pivoteo parcial**, o *pivoteo de columna máxima*, y se describe detalladamente en el algoritmo 6.2. En el algoritmo el intercambio real de renglones se simula intercambiando los valores de *NROW* en el paso 5.

ALGORITMO 6.2

Eliminación gaussiana con pivoteo parcial

Para resolver el sistema lineal $n \times n$

$$\begin{aligned} E_1: &\quad a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = a_{1,n+1} \\ E_2: &\quad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = a_{2,n+1} \\ &\quad \vdots \\ E_n: &\quad a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = a_{n,n+1} \end{aligned}$$

ENTRADA número de incógnitas y ecuaciones n; matriz aumentada $A = (a_{ij})$ donde $1 \le i \le n$ y $1 \le j \le n+1$.

SALIDA solución $x_1, \ldots, x_n$ o mensaje de que el sistema lineal no tiene una solución única.

Paso 1 Para $i = 1, \ldots, n$ tome $NROW(i) = i$. (*Inicializar apuntador de renglón.*)

Paso 2 Para $i = 1, \ldots, n-1$ haga los pasos 3-6. (*Proceso de eliminación.*)

Paso 3 Sea p el entero más pequeño con $i \le p \le n$ y $|a(NROW(p), i)| = \max_{i \le j \le n} |a(NROW(j), i)|$. (*Notación:* $a(NROW(i), j) \equiv a_{NROW_i, j}$.)

Paso 4 Si $a(NROW(p), i) = 0$ entonces SALIDA ('No existe una solución única'); PARAR.

Paso 5 Si $NROW(i) \ne NROW(p)$ entonces tome $NCOPY = NROW(i)$;
$NROW(i) = NROW(p)$;
$NROW(p) = NCOPY$.

(*Intercambio de renglones simulado.*)

Paso 6 Para $j = i + 1, \ldots, n$ haga los pasos 7 y 8.

Paso 7 Tome $m(NROW(j), i) = a(NROW(j), i)/ a(NROW(i), i)$.

Paso 8 Realice $(E_{NROW(j)} - m(NROW(j), i) \cdot E_{NROW(i)}) \to (E_{NROW(j)})$.

Paso 9 Si $a(NROW(n), n) = 0$ entonces SALIDA ('no existe una solución única'); PARAR.

Paso 10 Tome $x_n = a(NROW(n), n + 1)/a(NROW(n), n)$.
(*Comience la sustitución hacia atrás.*)

Paso 11 Para $i = n - 1, \ldots, 1$

$$\text{Tome } x_i = \frac{a(NROW(i), n+1) - \sum_{j=i+1}^{n} a(NROW(i), j) \cdot x_j}{a(NROW(i), i)}.$$

Paso 12 SALIDA $(x_1, \ldots, x_n)$; (*Procedimiento terminado exitosamente.*)
PARAR. ■

Cada multiplicador m_{ji} del algoritmo de pivoteo parcial tiene una magnitud menor o igual que 1. Aunque en la generalidad de los sistemas lineales esta estrategia es eficiente, se presentan situaciones donde resulta inadecuada.

EJEMPLO 3 El sistema lineal

$$\begin{aligned} E_1: &\quad 30.00x_1 + 591400x_2 = 591700, \\ E_2: &\quad 5.291x_1 - \quad 6.130x_2 = 46.78, \end{aligned}$$

es el mismo que el de los ejemplos 1 y 2, salvo que todos los elementos de la primera ecuación han sido multiplicados por 10^4. Con la aritmética de cuatro dígitos, el procedimiento descrito en el algoritmo 6.2 produciría los mismos resultados que se obtuvieron en el ejemplo 1. El valor máximo de la primera columna es 30.00 y el multiplicador

$$m_{21} = \frac{5.291}{30.00} = 0.1764$$

da origen al sistema

$$\begin{aligned} 30.00x_1 + 591400x_2 &\approx 591700, \\ -104300x_2 &\approx 104400, \end{aligned}$$

que tiene las mismas soluciones inexactas que el ejemplo 1: $x_2 \approx 1.001$ y $x_1 \approx -10.00$. ■

El pivoteo **parcial escalado**, denominado también *pivoteo de escalado de columna*, es el adecuado para el sistema del ejemplo 3. Coloca el elemento en el lugar del pivote más grande en relación con los elementos de su renglón. El primer paso del procedimiento consiste en definir, para cada renglón, un factor escalar s_i por medio de

$$s_i = \max_{1 \le j \le n} |a_{ij}|.$$

Si para alguna i tenemos $s_i = 0$, entonces el sistema no tiene una solución única, porque todos los elementos de i-ésimo renglón son cero. Suponiendo que no sea así, el intercam-

bio adecuado de renglones para poner ceros en la primera columna se determina seleccionando el menor entero p con

$$\frac{|a_{p1}|}{s_p} = \max_{1 \le k \le n} \frac{|a_{k1}|}{s_k}$$

y realizando $(E_1) \leftrightarrow (E_p)$. El cambio de escala garantiza que el mayor elemento de cada renglón tiene una magnitud *relativa* de 1 antes de realizar la comparación para el intercambio de renglones.

De manera análoga, antes de eliminar la variable x_i mediante las operaciones

$$E_k - m_{ki} E_i, \quad \text{para} \quad k = i + 1, \ldots, n,$$

elegimos el menor entero $p \ge i$ tal que

$$\frac{|a_{pi}|}{s_p} = \max_{i \le k \le n} \frac{|a_{ki}|}{s_k}$$

y realizamos el intercambio de renglones $E_i \leftrightarrow E_p$ si $i \ne p$. Los factores del cambio de escala $s_1, \ldots, s_n$ se calculan sólo una vez, al inicio del procedimiento y también deben intercambiarse al realizar los intercambios de renglones.

Al aplicar el pivoteo parcial escalado en el ejemplo 3, obtenemos

$$s_1 = \max\{|30.00|, |591400|\} = 591400$$

y

$$s_2 = \max\{|5.291|, |-6.130|\} = 6.130.$$

En consecuencia,

$$\frac{|a_{11}|}{s_1} = \frac{30.00}{591400} = 0.5073 \times 10^{-4}, \qquad \frac{|a_{21}|}{s_2} = \frac{5.291}{6.130} = 0.8631,$$

y se lleva a cabo el intercambio $(E_1) \leftrightarrow (E_2)$.

Al aplicar la eliminación gaussiana al nuevo sistema

$$5.291x_1 - 6.130x_2 = 46.78$$
$$30.00x_1 + 591400x_2 = 591700$$

obtenemos los resultados correctos $x_1 = 10.00$ y $x_2 = 1.000$.

En el algoritmo 6.3 se pone en ejecución el pivoteo parcial escalado.

ALGORITMO 6.3

Eliminación gaussiana con pivoteo parcial escalado

Los únicos pasos en este algoritmo que difieren de los del algoritmo 6.2 son:

Paso 1 Para $i = 1, \ldots, n$ tome $s_i = \max_{1 \le j \le n} |a_{ij}|$;
si $s_i = 0$ entonces SALIDA ('no existe una solución única')
PARAR.
tome $NROW(i) = i$.

Paso 2 Para $i = 1, \ldots, n - 1$ haga los pasos 3-6. (*Proceso de eliminación.*)

Paso 3 Sea p el entero más pequeño con $i \leq p \leq n$ y

$$\frac{|a(NROW(p), i)|}{s(NROW(p))} = \max_{i \leq j \leq n} \frac{|a(NROW(j), i)|}{s(NROW(j))}.$$ ■

En el ejemplo siguiente se explica el pivote parcial escalado por medio de Maple, con aritmética de redondeo a cifras finitas.

EJEMPLO 4 Resuelva el sistema lineal por medio de la aritmética de redondeos a tres dígitos.

$$\begin{aligned} 2.11x_1 - 4.21x_2 + 0.921x_3 &= 2.01, \\ 4.01x_1 + 10.2x_2 - 1.12x_3 &= -3.09, \\ 1.09x_1 + 0.987x_2 + 0.832x_3 &= 4.21. \end{aligned}$$

Para llamar a la aritmética de redondeo a tres dígitos, introducimos

```
>Digits:=3;
```

Tenemos $s_1 = 4.21$, $s_2 = 10.2$ y $s_3 = 1.09$. Por tanto,

$$\frac{|a_{11}|}{s_1} = \frac{2.11}{4.21} = 0.501, \quad \frac{|a_{21}|}{s_1} = \frac{4.01}{10.2} = 0.393 \quad \text{y} \quad \frac{|a_{31}|}{s_3} = \frac{1.09}{1.09} = 1.$$

La matriz aumentada AA se define mediante

```
>AA:=matriz(3,4,[2.11,-4.21,0.921,2.01,4.01,10.2,-1.12,-3.09,1.09,
                 0.987,0.832,4.21]);
```

que nos da

$$AA: = \begin{bmatrix} 2.11 & -4.21 & .921 & 2.01 \\ 4.01 & 10.2 & -1.12 & -3.09 \\ 1.09 & .987 & .832 & 4.21 \end{bmatrix}.$$

Puesto que $|a_{31}|/s_3$ es muy grande, efectuamos $(E_1) \leftrightarrow (E_3)$ usando

```
>AA:=swaprow(AA,1,3);
```

para obtener

$$AA: = \begin{bmatrix} 1.09 & .987 & .832 & 4.21 \\ 4.01 & 10.2 & -1.12 & -3.09 \\ 2.11 & -4.21 & .921 & 2.01 \end{bmatrix}.$$

Al calcular los multiplicadores obtenemos

```
>m21:=4.01/1.09;
```

$$m21: = 3.68$$

```
>m31:=2.11/1.09;
```

$$m31: = 1.94$$

Efectuamos las dos primeras eliminaciones por medio de

```
>AA:=addrow(AA,1,2,-m21);
```

y

```
>AA:=addrow(AA,1,3,-m31);
```

para obtener

$$AA: = \begin{bmatrix} 1.09 & .987 & .832 & 4.21 \\ 0 & 6.57 & -4.18 & -18.6 \\ 0 & -6.12 & -.689 & -6.16 \end{bmatrix}.$$

Dado que

$$\frac{|a_{22}|}{s_2} = \frac{6.57}{10.2} = 0.644 < \frac{|a_{32}|}{s_3} = \frac{6.12}{4.21} = 1.45,$$

efectuamos

```
>AA:=swaprow(AA,2,3);
```

lo cual nos da

$$AA: = \begin{bmatrix} 1.09 & .987 & .832 & 4.21 \\ 0 & -6.12 & -.689 & -6.16 \\ 0 & 6.57 & -4.18 & -18.6 \end{bmatrix}.$$

El multiplicador m_{32} se calcula mediante

```
>m32:=6.57/(-6.12);
```

$$m_{32} := -1.07.$$

La eliminación del paso

```
>AA:=addrow(AA,2,3, -m32);
```

nos da

$$AA: = \begin{bmatrix} 1.09 & .987 & .832 & 4.21 \\ 0 & -6.12 & -.689 & -6.16 \\ 0 & .02 & -4.92 & -25.2 \end{bmatrix}.$$

No podemos usar `backsub` por el elemento .02 en el lugar (3, 2). Este elemento es distinto de cero debido al error de redondeo, pero este pequeño problema se puede superar aplicando el comando

```
>AA[3,2]:=0;
```

que reemplaza al elemento .02 por 0. Para ver esto introducimos

```
>evalm(AA);
```

que muestra la matriz AA. Finalmente,

```
>x:=backsub(AA);
```

nos da la solución

$$x := [-.431 \quad .430 \quad 5.12].$$ ■

Los primeros cálculos adicionales que se requieren en el pivoteo parcial escalado resultan de determinar los factores escalares; en cada uno de los n renglones hay $(n-1)$ comparaciones, lo cual nos da un total de

$$n(n-1) \text{ comparaciones.}$$

Para determinar el primer cambio correcto, efectuamos n divisiones y luego $n-1$ comparaciones. Así pues, la primera determinación de intercambio agrega

$$n \text{ divisiones y } (n-1) \text{ comparaciones.}$$

Los factores de escala se calculan sólo una vez y, por lo mismo, el segundo paso requiere

$$(n-1) \text{ divisiones y } (n-2) \text{ comparaciones.}$$

Procederemos de manera similar hasta que haya ceros debajo de la diagonal principal de todos los renglones, menos del n-ésimo. En el caso final hay que realizar

$$2 \text{ divisiones y } 1 \text{ comparación.}$$

En consecuencia, el pivoteo parcial escalado agrega un total de

$$n(n-1) + \sum_{k=1}^{n-1} k = n(n-1) + \frac{(n-1)n}{2} = \frac{3}{2}n(n-1) \text{ comparaciones} \qquad (6.7)$$

y

$$\sum_{k=2}^{n} k = \sum_{k=1}^{n} k - 1 = \frac{n(n+1)}{2} - 1 \text{ divisiones}$$

al método de eliminación gaussiana. El tiempo que se requiere para efectuar una comparación es aproximadamente igual al que se requiere para efectuar una suma/resta. El tiempo total necesario para realizar el procedimiento básico de la eliminación gaussiana es $O(n^3/3)$ multiplicaciones/divisiones y $O(n^3/3)$ sumas/restas; por ello, con el pivoteo parcial escalado el tiempo de cálculo que se requiere para resolver un sistema de valores grandes de n no aumenta significativamente.

Para subrayar la importancia de seleccionar una sola vez los factores escalares, consideremos los cálculos adicionales que se necesitarían en caso de modificar el procedimien-

to, de modo que los nuevos factores escalares se determinarán cada vez que se tome una decisión de intercambio de renglones. En este caso, el término $n(n-1)$ de la ecuación (6.7) se reemplazará por

$$\sum_{k=2}^{n} k(k-1) = \frac{1}{3}n(n^2-1).$$

En consecuencia, este método de pivoteo agregaría $O(n^3/3)$ comparaciones, además de las $[n(n+1)/2]-1$ divisiones. Si un sistema exige este tipo de pivoteo, se usará a cambio **el pivoteo completo** (o *máximo*). El pivoteo completo en el k-ésimo paso busca todos los elementos a_{ij}, para $i = k, k+1, \ldots, n$ y para $j = k, k+1, \ldots, n$, a fin de localizar el de mayor magnitud. Los intercambios de renglones y columnas se realizan para poner este elemento en la posición de pivote. El primer paso del pivoteo total requiere hacer n^2-1 comparaciones, el segundo requiere hacer $(n-1)^2-1$ comparaciones y así sucesivamente. Así pues, el tiempo adicional necesario para incorporar el pivoteo completo en la eliminación gaussiana es

$$\sum_{k=2}^{n}(k^2-1) = \frac{n(n-1)(2n+5)}{6}$$

comparaciones. La cifra anterior se parece al número que requiere el método modificado de pivoteo escalado de columna, pero no se necesitan divisiones. En consecuencia, el pivoteo completo es la estrategia que se recomienda para los sistemas donde la exactitud es indispensable y puede justificarse el tiempo de ejecución que se requiere.

CONJUNTO DE EJERCICIOS 6.2

1. Obtenga los intercambios de renglones que se requieren para resolver los siguientes sistemas lineales, utilizando para ello el algoritmo 6.1.

   a. $x_1 - 5x_2 + x_3 = 7,$
   $10x_1 + 20x_3 = 6,$
   $5x_1 - x_3 = 4.$

   b. $x_1 + x_2 - x_3 = 1,$
   $x_1 + x_2 + 4x_3 = 2,$
   $2x_1 - x_2 + 2x_3 = 3.$

   c. $2x_1 - 3x_2 + 2x_3 = 5,$
   $-4x_1 + 2x_2 - 6x_3 = 14,$
   $2x_1 + 2x_2 + 4x_3 = 8.$

   d. $x_2 + x_3 = 6,$
   $x_1 - 2x_2 - x_3 = 4,$
   $x_1 - x_2 + x_3 = 5.$

2. Repita el ejercicio 1 usando el algoritmo 6.2.
3. Repita el ejercicio 1 usando el algoritmo 6.3.
4. Repita el ejercicio 1 usando el pivoteo completo.
5. Use la eliminación gaussiana y la aritmética de truncamiento de tres dígitos para resolver los siguientes sistemas lineales. Después, compare las aproximaciones con la solución real.

   a. $0.03x_1 + 58.9x_2 = 59.2$
   $5.31x_1 - 6.10x_2 = 47.0.$
   Solución real $(10, 1)^t$.

   b. $58.9x_1 + 0.03x_2 = 59.2,$
   $-6.10x_1 + 5.31x_2 = 47.0.$
   Solución real $(1, 10)^t$.

c.
$$\begin{aligned} 3.03x_1 - 12.1x_2 + 14x_3 &= -119, \\ -3.03x_1 + 12.1x_2 - 7x_3 &= 120, \\ 6.11x_1 - 14.2x_2 + 21x_3 &= -139. \end{aligned}$$

Solución real $(0, 10, \frac{1}{7})^t$.

d.
$$\begin{aligned} 3.3330x_1 + 15920x_2 + 10.333x_3 &= 7953, \\ 2.2220x_1 + 16.710x_2 + 9.6120x_3 &= 0.965, \\ -1.5611x_1 + 5.1792x_2 - 1.6855x_3 &= 2.714. \end{aligned}$$

Solución real $(1, 0.5, -1)^t$.

e.
$$\begin{aligned} 1.19x_1 + 2.11x_2 - 100x_3 + x_4 &= 1.12, \\ 14.2x_1 - 0.122x_2 + 12.2x_3 - x_4 &= 3.44, \\ 100x_2 - 99.9x_3 + x_4 &= 2.15, \\ 15.3x_1 + 0.110x_2 - 13.1x_3 - x_4 &= 4.16. \end{aligned}$$

Solución real $(0.17682530, 0.01269269, -0.02065405, -1.18260870)^t$.

f.
$$\begin{aligned} \pi x_1 - e x_2 + \sqrt{2} x_3 - \sqrt{3} x_4 &= \sqrt{11}, \\ \pi^2 x_1 + e x_2 - e^2 x_3 + \tfrac{3}{7} x_4 &= 0, \\ \sqrt{5} x_1 - \sqrt{6} x_2 + x_3 - \sqrt{2} x_4 &= \pi, \\ \pi^3 x_1 + e^2 x_2 - \sqrt{7} x_3 + \tfrac{1}{9} x_4 &= \sqrt{2}. \end{aligned}$$

Solución real $(0.78839378, -3.12541367, 0.16759660, 4.55700252)^t$.

6. Repita el ejercicio 5 usando la aritmética de redondeo a tres dígitos.

7. Repita el ejercicio 5 usando la eliminación gaussiana con pivoteo parcial.

8. Repita el ejercicio 6 usando la eliminación gaussiana con pivoteo parcial.

9. Repita el ejercicio 5 usando la eliminación gaussiana con pivoteo parcial escalado.

10. Repita el ejercicio 6 usando la eliminación gaussiana con pivoteo parcial escalado.

11. Repita el ejercicio 5 usando el algoritmo 6.1 con la aritmética computacional de precisión simple.

12. Repita el ejercicio 5 usando el algoritmo 6.2 con la aritmética computacional de precisión simple.

13. Repita el ejercicio 5 usando el algoritmo 6.3 con la aritmética computacional de precisión simple.

14. Construya un algoritmo para el procedimiento de pivoteo completo que se explica en el texto.

15. Use el algoritmo de pivoteo completo que se explica en el ejercicio 14 para obtener las soluciones de

a. Ejercicio 5 **b.** Ejercicio 6 **c.** Ejercicio 11

16. Suponga que

$$\begin{aligned} 2x_1 + x_2 + 3x_3 &= 1, \\ 4x_1 + 6x_2 + 8x_3 &= 5, \\ 6x_1 + \alpha x_2 + 10x_3 &= 5, \end{aligned}$$

con $|\alpha| < 10$. ¿Para cuál de los siguientes valores de α no se requerirá intercambio de renglones al resolver este sistema mediante el pivoteo parcial escalado?

a. $\alpha = 6$ **b.** $\alpha = 9$ **c.** $\alpha = -3$

6.3 Álgebra lineal e inversas de matrices

En la sección 6.1 dijimos que las matrices constituyen un método adecuado para expresar y tratar un sistema lineal. En esta sección estudiaremos un poco del álgebra asociada a ellas y explicaremos cómo utilizarla para resolver problemas que contengan sistemas lineales.

Definición 6.2 Dos matrices A y B son **iguales** si tienen el mismo tamaño, digamos $n \times m$, y si $a_{ij} = b_{ij}$ para cada $i = 1, 2, \ldots, n$ y para cada $j = 1, 2, \ldots, m$ ■

Esta definición implica, por ejemplo, que

$$\begin{bmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{bmatrix} \neq \begin{bmatrix} 2 & 3 \\ -1 & 1 \\ 7 & 0 \end{bmatrix},$$

porque su dimensión es diferente.

Dos operaciones importantes que se realizan con matrices son la suma de dos matrices y de multiplicación de una matriz por un número real.

Definición 6.3 Si A y B son ambas matrices de $n \times m$, entonces la **suma** de A y B, denotada $A + B$, es la matriz de $n \times m$ cuyos elementos son $a_{ij} + b_{ij}$, para cada $i = 1, 2, \ldots, n$ y para cada $j = 1, 2, \ldots, m$. ■

Definición 6.4 Si A es una matriz de $n \times m$ y si λ es un número real, entonces la **multiplicación por escalar** de λ y A, denotada λA, es la matriz de $n \times m$ cuyos elementos son λa_{ij}, para cada $i = 1, 2, \ldots, n$ y para cada $j = 1, 2, \ldots, m$. ■

EJEMPLO 1 Sea

$$A = \begin{bmatrix} 2 & -1 & 7 \\ 3 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 2 & -8 \\ 0 & 1 & 6 \end{bmatrix}$$

y $\lambda = -2$. Entonces

$$A + B = \begin{bmatrix} 2+4 & -1+2 & 7-8 \\ 3+0 & 1+1 & 0+6 \end{bmatrix} = \begin{bmatrix} 6 & 1 & -1 \\ 3 & 2 & 6 \end{bmatrix}$$

y

$$\lambda A = \begin{bmatrix} -2(2) & -2(-1) & -2(7) \\ -2(3) & -2(1) & -2(0) \end{bmatrix} = \begin{bmatrix} -4 & 2 & -14 \\ -6 & -2 & 0 \end{bmatrix}.$$ ■

Sea O una matriz cuyos elementos son todos cero y sea $-A$ la matriz cuyos elementos son $-a_{ij}$. En la suma de matrices y en la multiplicación por escalar tenemos las siguientes propiedades generales. Dichas propiedades son suficientes para clasificar el conjunto de todas las matrices de $n \times m$ con elementos reales como un **espacio vectorial** sobre el campo de los números reales. (Véase [ND, pp. 107-109].)

Teorema 6.5 Sean A, B y C matrices de $n \times m$ y sean λ y μ números reales. Se aplican las siguientes propiedades de suma y multiplicación por escalar:

a. $A + B = B + A$,
b. $(A + B) + C = A + (B + C)$,
c. $A + 0 = 0 + A = A$,
d. $A + (-A) = -A + A = 0$,
e. $\lambda(A + B) = \lambda A + \lambda B$,
f. $(\lambda + \mu)A = \lambda A + \mu A$,
g. $\lambda(\mu A) = (\lambda\mu)A$,
h. $1A = A$.

Las propiedades anteriores se deducen de resultados similares acerca de los números reales. ■

Definición 6.6 Sea A una matriz de $n \times m$ y sea B una matriz de $m \times p$. El **producto matricial** de A y B, denotado AB, es una matriz C de $n \times p$ cuyos elementos c_{ij} están dados por

$$c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj}$$

para cada $i = 1, 2, \ldots, n$ y para cada $j = 1, 2, \ldots, p$. ■

Podemos considerar el cálculo de c_{ij} como la multiplicación de elementos del i-ésimo renglón de A con los elementos correspondientes de la j-ésima columna de B, seguida de la suma. Es decir,

$$[a_{i1}, a_{i2}, \ldots, a_{im}] \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{mj} \end{bmatrix} = c_{ij},$$

así que

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj} = \sum_{k=1}^{m} a_{ik}b_{kj}.$$

Esto explica por qué en el producto AB por definir, el número de columnas de A debe ser igual al número de renglones de B.

El siguiente ejemplo servirá para hacer aún más claro el proceso de multiplicación de matrices.

EJEMPLO 2 Sean

$$A = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 2 \\ 0 & -2 & -3 \end{bmatrix}, \qquad B = \begin{bmatrix} 3 & 2 \\ -1 & 1 \\ 6 & 4 \end{bmatrix},$$

$$C = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 3 & 2 \end{bmatrix} \quad \text{y} \quad D = \begin{bmatrix} 1 & -1 & 1 \\ 2 & -1 & 2 \\ 3 & 0 & 3 \end{bmatrix}.$$

Entonces,

$$AD = \begin{bmatrix} 1 & -3 & 1 \\ -11 & -4 & 11 \\ -13 & 2 & -13 \end{bmatrix} \neq \begin{bmatrix} -1 & -2 & -6 \\ 1 & -3 & -10 \\ 6 & -3 & -12 \end{bmatrix} = DA.$$

Además,

$$BC = \begin{bmatrix} 4 & 9 & 4 \\ -3 & 2 & 2 \\ 8 & 18 & 8 \end{bmatrix} \quad \text{y} \quad CB = \begin{bmatrix} 5 & 5 \\ 6 & 9 \end{bmatrix}$$

ni siquiera son del mismo tamaño.

Finalmente,

$$AB = \begin{bmatrix} -1 & 1 \\ 20 & 15 \\ -16 & -14 \end{bmatrix},$$

pero no es posible calcular BA. ■

Definición 6.7 Una matriz **cuadrada** tiene la misma cantidad de renglones que de columnas. Una matriz **diagonal** es una matriz cuadrada $D = (d_{ij})$ con $d_{ij} = 0$ siempre que $i \neq j$. La **matriz identidad de orden** n, $I_n = (\delta_{ij})$, es una matriz diagonal con los elementos

$$\delta_{ij} = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j. \end{cases}$$

Cuando el tamaño de I_n está claro, esta matriz suele escribirse simplemente como I. ■

Por ejemplo, la matriz identidad de orden tres es

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Definición 6.8 Una matriz **triangular superior** de $n \times n$ $U = (u_{ij})$ tiene, para toda $j = 1, 2, \ldots, n$, los elementos

$$u_{ij} = 0, \text{ para cada } i = j+1, j+2, \ldots, n;$$

y una matriz **triangular inferior** $L = (l_{ij})$ tiene, para toda $j = 1, 2, \ldots, n$, los elementos

$$l_{ij} = 0, \text{ para cada } i = 1, 2, \ldots, j-1.$$ ■

EJEMPLO 3 La matriz identidad de orden tres es

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Si A es una matriz cualquiera de 3×3, entonces

$$AI_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = A. \quad \blacksquare$$

La matriz identidad I_n conmuta con cualquier matriz A de $n \times n$, es decir, no importa el orden de la multiplicación $I_nA = A = AI_n$. En el ejemplo 2 vimos que la propiedad $AB = BA$ no suele ser verdadera en la multiplicación de matrices. En el siguiente teorema veremos algunas de las propiedades referentes a la multiplicación de matrices.

Teorema 6.9 Sea A una matriz de $n \times m$, B una matriz de $m \times k$, C una matriz de $k \times p$, D una matriz de $m \times k$, y λ un número real. Se aplican las siguientes propiedades:

a. $A(BC) = (AB)C$;

b. $A(B + D) = AB + AD$;

c. $I_mB = B$ y $BI_k = B$;

d. $\lambda(AB) = (\lambda A)B = A(\lambda B)$.

Demostración Para mostrar el método en cuestión, sólo se incluye la verificación de la propiedad del inciso (a). Las otras partes pueden resolverse de modo parecido.

Para demostrar que $A(BC) = (AB)C$, calcule la entrada i, j de ambos lados de la ecuación. BC es una matriz de $m \times p$ con entrada i, j.

$$(BC)_{sj} = \sum_{l=1}^{k} b_{sl}c_{lj}.$$

Así pues, $A(BC)$ es una matriz de $n \times p$ con entradas

$$[A(BC)]_{ij} = \sum_{s=1}^{k} a_{is}(BC)_{sj} = \sum_{s=1}^{k} a_{is} \left(\sum_{l=1}^{k} b_{sl}c_{lj} \right) = \sum_{s=1}^{k} \sum_{l=1}^{k} a_{is}b_{sl}c_{lj}.$$

De manera análoga, AB es una matriz de $n \times k$ con entradas

$$(AB)_{il} = \sum_{s=1}^{k} a_{is}b_{sl},$$

por tanto, (AB) es una matriz de $n \times p$ con entradas

$$[(AB)C]_{ij} = \sum_{l=1}^{k} (AB)_{il}c_{lj} = \sum_{l=1}^{k} \left(\sum_{s=1}^{k} a_{is}b_{sl} \right) c_{lj} = \sum_{l=1}^{k} \sum_{s=1}^{k} a_{is}b_{sl}c_{lj}.$$

Al intercambiar el orden de la suma en el lado derecho, obtenemos

$$[(AB)C]_{ij} = \sum_{s=1}^{k} \sum_{l=1}^{k} a_{is}b_{sl}c_{lj} = [A(BC)]_{ij},$$

para toda $i = 1, 2, \ldots, n$ y para toda $j = 1, 2, \ldots, p$. Por tanto, $A(BC) = (AB)C$. ■ ■ ■

Podemos considerar el sistema lineal

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

como la ecuación matricial

$$A\mathbf{x} = \mathbf{b},$$

donde

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

La **inversa de una matriz** está relacionada con los sistemas lineales.

Definición 6.10 Se dice que una matriz A de $n \times n$ es **no singular**, si existe una matriz A^{-1} de $n \times n$ con $AA^{-1} = A^{-1}A = I$. La matriz A^{-1} se llama **inversa de** A. Una matriz que no tiene inversa se le da el nombre de **singular**. ■

Las siguientes propiedades referentes a las inversas de matrices se deducen de la definición 6.10. Las demostraciones de estos resultados se incluyen en el ejercicio 5.

Teorema 6.11 Para una matriz A no singular de $n \times n$:

a. A^{-1} es única.

b. A^{-1} es no singular y $(A^{-1})^{-1} = A$.

c. Si B también es una matriz no singular de $n \times n$, entonces

$$(AB)^{-1} = B^{-1}A^{-1}.$$ ■

EJEMPLO 4 Sean

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Entonces

$$AB = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_3.$$

De modo semejante, $BA = I_3$, por tanto, A y B son no singulares con $B = A^{-1}$ y $A = B^{-1}$.

Si tenemos la inversa de A, será fácil resolver un sistema lineal de la forma $A\mathbf{x} = \mathbf{b}$. Supongamos, por ejemplo, que queremos resolver

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 2, \\ 2x_1 + x_2 \phantom{{}-2x_3} &= 3, \\ -x_1 + x_2 + 2x_3 &= 4. \end{aligned}$$

Primero, convertimos el sistema en la ecuación matricial

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix},$$

y luego multiplicamos ambos lados por la inversa:

$$\begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \left(\begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right) = \begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{7}{9} \\ \frac{13}{9} \\ \frac{5}{3} \end{bmatrix},$$

así que

$$\begin{bmatrix} \frac{7}{9} \\ \frac{13}{9} \\ \frac{5}{3} \end{bmatrix} = \left(\begin{bmatrix} -\frac{2}{9} & \frac{5}{9} & -\frac{1}{9} \\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{3}{9} & \frac{3}{9} & \frac{3}{9} \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = I_3 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Esto nos da la solución $x_1 = 7/9$, $x_2 = 13/9$ y $x_3 = 5/3$. ■

Aunque es fácil resolver un sistema lineal de la forma $A\mathbf{x} = \mathbf{b}$ si se conoce A^{-1}, desde el punto de vista de los cálculos necesarios, no es eficiente determinar A^{-1} a fin de resolver el sistema. (Véase el ejercicio 8.) Pese a ello, desde el punto de vista conceptual conviene describir un método que determine la inversa de una matriz.

A fin de encontrar un método para calcular A^{-1} suponiendo su existencia, consideremos nuevamente la multiplicación de matrices. Sea B_j la j-ésima columna de la matriz B de $n \times n$,

$$B_j = \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix}.$$

Si $AB = C$, entonces la j-ésima columna de C está dada por el producto

$$\begin{bmatrix} c_{1j} \\ c_{2j} \\ \vdots \\ c_{nj} \end{bmatrix} = C_j = AB_j = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{nj} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} a_{1k}b_{kj} \\ \sum_{k=1}^{n} a_{2k}b_{kj} \\ \vdots \\ \sum_{k=1}^{n} a_{nk}b_{kj} \end{bmatrix}.$$

Supongamos que A^{-1} existe y que $A^{-1} = B = (b_{ij})$. Entonces $AB = I$ y

$$AB_j = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{donde el valor 1 aparece en el } j\text{-ésimo renglón.}$$

Para encontrar B debemos resolver n sistemas lineales donde la j-ésima columna de la inversa es la solución del sistema lineal que en el lado derecho tiene la j-ésima columna de I. En el siguiente ejemplo se demuestra el método en cuestión.

EJEMPLO 5 Para determinar la inversa de la matriz

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix},$$

consideremos primero el producto AB, donde B es una matriz arbitraria de 3×3.

$$AB = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

$$= \begin{bmatrix} b_{11} + 2b_{21} - b_{31} & b_{12} + 2b_{22} - b_{32} & b_{13} + 2b_{23} - b_{33} \\ 2b_{11} + b_{21} & 2b_{12} + b_{22} & 2b_{13} + b_{23} \\ -b_{11} + b_{21} + 2b_{31} & -b_{12} + b_{22} + 2b_{32} & -b_{13} + b_{23} + 2b_{33} \end{bmatrix}.$$

Si $B = A^{-1}$, entonces $AB = I$, de modo que tenemos

$$\begin{aligned} b_{11} + 2b_{21} - b_{31} &= 1, \\ 2b_{11} + b_{21} &= 0, \\ -b_{11} + b_{21} + 2b_{31} &= 0, \end{aligned} \qquad \begin{aligned} b_{12} + 2b_{22} - b_{32} &= 0, \\ 2b_{12} + b_{22} &= 1, \\ -b_{12} + b_{22} + 2b_{32} &= 0, \end{aligned} \qquad \begin{aligned} b_{13} + 2b_{23} - b_{33} &= 0, \\ 2b_{13} + b_{23} &= 0, \\ -b_{13} + b_{23} + 2b_{33} &= 1. \end{aligned}$$

Nótese que los coeficientes de los sistemas de ecuaciones son los mismos; el único cambio ocurre en el lado derecho de las ecuaciones. En consecuencia, podemos realizar la eliminación gaussiana en una matriz aumentada más grande, que se forma al combinar las matrices de los sistemas:

$$\left[\begin{array}{rrr|rrr} 1 & 2 & -1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 2 & 0 & 0 & 1 \end{array}\right].$$

Primero $(E_2 - 2E_1) \to (E_2)$ y $(E_3 + E_1) \to (E_3)$ luego $(E_3 + E_2) \to (E_3)$ producen

$$\left[\begin{array}{rrr|rrr} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -3 & 2 & -2 & 1 & 0 \\ 0 & 3 & 1 & 1 & 0 & 1 \end{array}\right] \quad \text{y} \quad \left[\begin{array}{rrr|rrr} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -3 & 2 & -2 & 1 & 0 \\ 0 & 0 & 3 & -1 & 1 & 1 \end{array}\right].$$

La sustitución hacia atrás se efectúa en cada una de las tres matrices aumentadas.

$$\left[\begin{array}{rrr:r}1 & 2 & -1 & 1\\0 & -3 & 2 & -2\\0 & 0 & 3 & -1\end{array}\right],\quad\left[\begin{array}{rrr:r}1 & 2 & -1 & 0\\0 & -3 & 2 & 1\\0 & 0 & 3 & 1\end{array}\right],\quad\left[\begin{array}{rrr:r}1 & 2 & -1 & 0\\0 & -3 & 2 & 0\\0 & 0 & 3 & 1\end{array}\right],$$

y obtenemos

$$\begin{array}{llll}b_{11} = -\frac{2}{9}, & b_{12} = \frac{5}{9}, & & b_{13} = -\frac{1}{9},\\ b_{21} = \frac{4}{9}, & b_{22} = -\frac{1}{9}, & \text{y} & b_{23} = \frac{2}{9},\\ b_{31} = -\frac{1}{3}, & b_{32} = \frac{1}{3}, & & b_{33} = \frac{1}{3}.\end{array}$$

Como se muestra en el ejemplo 4, estos son los elementos de A^{-1}:

$$A^{-1} = \begin{bmatrix}-\frac{2}{9} & \frac{5}{9} & -\frac{1}{9}\\ \frac{4}{9} & -\frac{1}{9} & \frac{2}{9}\\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}$$ ■

En el último ejemplo explicamos el cálculo de A^{-1}. Como vimos en ese ejemplo, conviene crear la matriz aumentada más grande,

$$[A \;\vdots\; I].$$

Luego de hacer las eliminación conforme al algoritmo 6.1, obtenemos una matriz aumentada de la forma

$$[U \;\vdots\; Y],$$

donde U es la matriz triangular superior y Y es la matriz que se obtiene al efectuar las mismas operaciones en la matriz identidad I que realizamos para llevar A en U.

La eliminación gaussiana con sustitución hacia atrás requiere $4n^3/3 - n/3$ multiplicaciones/divisiones y $4n^3/3 - 3n^2/2 + n/6$ sumas/restas para resolver los n sistemas lineales (véase el ejercicio 8(a)). Hay que tener mucho cuidado en la ejecución y señalar las operaciones que *no* deben realizarse, como por ejemplo, una multiplicación cuando se sabe que uno de los multiplicadores es la unidad o una resta cuando se sabe que el sustraendo es cero. Entonces, el número necesario de multiplicaciones/divisiones podrá reducirse a n^3 y la cantidad de sumas/restas podrá reducirse a $n^3 - 2n^2 + n$ (véase el ejercicio 8(d)).

Otra matriz importante asociada a una matriz A dada es la *transpuesta*, denotada A^t.

Definición 6.12 La **transpuesta** de una matriz A de $n \times m$ $A = (a_{ij})$ es una matriz A^t, donde para cada i, la i-ésima columna de A^t es la misma que el i-ésimo renglón de A, es decir, $A^t = (a_{ji})$. Una matriz cuadrada A será **simétrica** si $A = A^t$. ■

Por ejemplo, las matrices

$$A = \begin{bmatrix}7 & 2 & 0\\3 & 5 & -1\\0 & 5 & -6\end{bmatrix},\quad B = \begin{bmatrix}2 & 4 & 7\\3 & -5 & -1\end{bmatrix},\quad C = \begin{bmatrix}6 & 2 & -3\\4 & -2 & 0\\-3 & 0 & 1\end{bmatrix}$$

tienen las respuestas

$$A^t = \begin{bmatrix} 7 & 3 & 0 \\ 2 & 5 & 5 \\ 0 & -1 & -6 \end{bmatrix}, \quad B^t = \begin{bmatrix} 2 & 3 \\ 4 & -5 \\ 7 & -1 \end{bmatrix}, \quad C^t = \begin{bmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{bmatrix}.$$

La matriz C es simétrica porque $C^t = C$, no así las matrices A y B.

La demostración del siguiente resultado se deduce directamente de la definición de transpuesta.

Teorema 6.13 Las siguientes operaciones relativas a la transpuesta de una matriz son válidas siempre que sean factibles:

a. $(A^t)^t = A$,

b. $(A + B)^t = A^t + B^t$,

c. $(AB)^t = B^t A^t$,

d. Si A^{-1} existe $(A^{-1})^t = (A^t)^{-1}$. ■

Podemos usar cualquier sistema de álgebra por computadora para efectuar estas operaciones aritméticas. La suma de matrices se realiza en Maple con `matadd(A,B)` o con `evalm(A+B)`. La multiplicación por escalar se define por medio de `scalarmul(A,c)` o bien `evalm(c*A)`.

La multiplicación de matrices se hace usando `multiply(A,B)` o `evalm(A&*B)`. La transposición de matrices se obtiene mediante `transpose(A)` y la inversa matricial por medio de `inverse(A)`.

CONJUNTO DE EJERCICIOS 6.3

1. Determine cuál de las siguientes matrices son no singulares y calcule su inversa.

a. $\begin{bmatrix} 4 & 2 & 6 \\ 3 & 0 & 7 \\ -2 & -1 & -3 \end{bmatrix}$

b. $\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & -1 \\ 3 & 1 & 1 \end{bmatrix}$

c. $\begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

d. $\begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & 2 & -4 & -2 \\ 2 & 1 & 1 & 5 \\ -1 & 0 & -2 & -4 \end{bmatrix}$

e. $\begin{bmatrix} 4 & 0 & 0 & 0 \\ 6 & 7 & 0 & 0 \\ 9 & 11 & 1 & 0 \\ 5 & 4 & 1 & 1 \end{bmatrix}$

f. $\begin{bmatrix} 2 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 2 & -1 & 3 & 1 \\ 3 & -1 & 4 & 3 \end{bmatrix}$

2. Considere los cuatro sistemas lineales de 3×3 que tienen la misma matriz de coeficientes.

$$\begin{aligned} 2x_1 - 3x_2 + x_3 &= 2, \\ x_1 + x_2 - x_3 &= -1, \\ -x_1 + x_2 - 3x_3 &= 0; \end{aligned} \qquad \begin{aligned} 2x_1 - 3x_2 + x_3 &= 6, \\ x_1 + x_2 - x_3 &= 4, \\ -x_1 + x_2 - 3x_3 &= 5; \end{aligned}$$

$$\begin{aligned} 2x_1 - 3x_2 + x_3 &= 0, \\ x_1 + x_2 - x_3 &= 1, \\ -x_1 + x_2 - 3x_3 &= -3; \end{aligned} \qquad \begin{aligned} 2x_1 - 3x_2 + x_3 &= -1, \\ x_1 + x_2 - x_3 &= 0, \\ -x_1 + x_2 - 3x_3 &= 0. \end{aligned}$$

a. Resuelva los sistemas lineales aplicando la eliminación gaussiana a la matriz aumentada.

$$\left[\begin{array}{rrr:rrrr} 2 & -3 & 1 & 2 & 6 & 0 & -1 \\ 1 & 1 & -1 & -1 & 4 & 1 & 0 \\ -1 & 1 & -3 & 0 & 5 & -3 & 0 \end{array}\right].$$

b. Resuelva los sistemas lineales encontrando y multiplicándolos por la inversa de

$$A = \begin{bmatrix} 2 & -3 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & -3 \end{bmatrix}.$$

c. ¿Cuál método requiere más operaciones?

3. Repita el ejercicio 2 usando los sistemas lineales

$$\begin{aligned} x_1 - x_2 + 2x_3 - x_4 &= 6, & x_1 - x_2 + 2x_3 - x_4 &= 1, \\ x_1 \quad\;\; - x_3 + x_4 &= 4, & x_1 \quad\;\; - x_3 + x_4 &= 1, \\ 2x_1 + x_2 + 3x_3 - 4x_4 &= -2, & 2x_1 + x_2 + 3x_3 - 4x_4 &= 2, \\ -x_2 + x_3 - x_4 &= 5; & -x_2 + x_3 - x_4 &= -1. \end{aligned}$$

4. Demuestre las siguientes afirmaciones o bien dé contraejemplos para probar que no son verdaderas.

a. El producto de dos matrices simétricas es simétrico.

b. La inversa de una matriz simétrica no singular es una matriz simétrica no singular.

c. Si A y B son matrices de $n \times n$, entonces $(AB)^t = A^tB^t$.

5. Se requieren las siguientes afirmaciones para demostrar el teorema 6.11.

a. Demuestre que, si existe A^{-1} es única.

b. Demuestre que, si A es no singular, entonces $(A^{-1})^{-1} = A$.

c. Demuestre que, si A y B son matrices no singulares de $n \times n$, entonces $(AB)^{-1} = B^{-1}A^{-1}$.

6. Demuestre el teorema 6.5.

7. a. Demuestre que el producto de dos matrices triangulares inferiores de $n \times n$ es triangular inferior.

b. Demuestre que el producto de dos matrices triangulares superiores de $n \times n$ es triangular superior.

c. Demuestre que la inversa de una matriz triangular inferior no singular de $n \times n$ es triangular inferior.

8. Suponga que debe resolver m sistemas lineales

$$A\mathbf{x}^{(p)} = \mathbf{b}^{(p)}, \quad p = 1, 2, \ldots, m,$$

que tienen las matrices de coeficientes A de $n \times n$.

a. Demuestre que la eliminación gaussiana con sustitución hacia atrás aplicada a la matriz aumentada

$$[A{:}\ \mathbf{b}^{(1)}\mathbf{b}^{(2)} \cdots \mathbf{b}^{(m)}]$$

requiere

$$\frac{1}{3}n^3 + mn^2 - \frac{1}{3}n \text{ multiplicaciones/divisiones}$$

y

$$\frac{1}{3}n^3 + mn^2 - \frac{1}{2}n^2 - mn + \frac{1}{6}n \text{ sumas/restas.}$$

b. Demuestre que el método de Gauss-Jordan (véase el ejercicio 8, sección 6.1) aplicado a la matriz aumentada,

$$[A: \quad \mathbf{b}^{(1)}\mathbf{b}^{(2)} \cdots \mathbf{b}^{(m)}]$$

requiere

$$\frac{1}{2}n^3 + mn^2 - \frac{1}{2}n \text{ multiplicaciones/divisiones}$$

y

$$\frac{1}{2}n^3 + (m-1)n^2 + \left(\frac{1}{2} - m\right)n \text{ sumas/restas.}$$

c. Para el caso especial

$$\mathbf{b}^{(p)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow p\text{-ésimo renglón,}$$

para toda $p = 1, \ldots, m$, con $m = n$, la solución $\mathbf{x}^{(p)}$ es la p-ésima columna de A^{-1}. Demuestre que la eliminación gaussiana hacia atrás requiere

$$\frac{4}{3}n^3 - \frac{1}{3}n \text{ multiplicaciones/divisiones}$$

y

$$\frac{4}{3}n^3 - \frac{3}{2}n^2 + \frac{1}{6}n \text{ sumas/restas}$$

para esta aplicación y que el método de Gauss-Jordan requiere

$$\frac{3}{2}n^3 - \frac{1}{2}n \text{ multiplicaciones/divisiones}$$

y

$$\frac{3}{2}n^3 - 2n^2 + \frac{1}{2}n \text{ sumas/restas.}$$

d. Use la eliminación gaussiana y construya un algoritmo para obtener A^{-1}, pero no realice las multiplicaciones cuando sepa que uno de los multiplicadores es la unidad y tampoco efectúe las sumas/restas cuando sepa que uno de los elementos es cero. Demuestre que los cálculos requeridos se reducen a n^3 multiplicaciones/divisiones y $n^3 - 2n^2 + n$ sumas/restas.

e. Demuestre que para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$, cuando se conoce A^{-1} requiere n^2 multiplicaciones/divisiones y $n^2 - n$ sumas/restas.

f. Demuestre que para resolver m sistemas lineales $A\mathbf{x}^{(p)} = \mathbf{b}^{(p)}$, para $p = 1, 2, \ldots, m$, mediante el método $\mathbf{x}^{(p)} = A^{-1}\mathbf{b}^{(p)}$ se requieren mn^2 multiplicaciones y $m(n^2 - n)$ sumas, si se conoce A^{-1}.

g. Sea A una matriz de $n \times n$. Compare el número de operaciones necesarias para resolver n sistemas lineales que contenga A, mediante la eliminación gaussiana con sustitución hacia atrás e invirtiendo primero A y multiplicando luego $A\mathbf{x} = \mathbf{b}$ por A^{-1}, para $n = 3, 10, 50, 100$. ¿Conviene calcular A^{-1} a fin de resolver los sistemas lineales?

9. Use el algoritmo que se desarrolló en el ejercicio 8(d) y obtenga con él las matrices no singulares del ejercicio 1.

10. A menudo conviene dividir las matrices en un conjunto de submatrices. Por ejemplo, podemos dividir las matrices

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 3 & -4 & -3 \\ 6 & 5 & 0 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 2 & -1 & 7 & 0 \\ 3 & 0 & 4 & 5 \\ -2 & 1 & -3 & 1 \end{bmatrix}$$

en

$$\left[\begin{array}{cc:c} 1 & 2 & -1 \\ 3 & -4 & -3 \\ \hdashline 6 & 5 & 0 \end{array}\right] = \left[\begin{array}{c:c} A_{11} & A_{12} \\ \hdashline A_{21} & A_{22} \end{array}\right]$$

y

$$\left[\begin{array}{ccc:c} 2 & -1 & 7 & 0 \\ 3 & 0 & 4 & 5 \\ \hdashline -2 & 1 & -3 & 1 \end{array}\right] = \left[\begin{array}{c:c} B_{11} & B_{12} \\ \hdashline B_{21} & B_{22} \end{array}\right]$$

a. Demuestre que en este caso el producto de A y B es

$$AB = \left[\begin{array}{c:c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hdashline A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{array}\right]$$

b. Si en lugar de lo anterior dividiéramos B en

$$B = \left[\begin{array}{ccc:c} 2 & -1 & 7 & 0 \\ \hdashline 3 & 0 & 4 & 5 \\ -2 & 1 & -3 & 1 \end{array}\right] = \left[\begin{array}{c:c} B_{11} & B_{12} \\ \hdashline B_{21} & B_{22} \end{array}\right],$$

¿sería válido el resultado del inciso (a)?

c. Haga una conjetura sobre las condiciones necesarias para que el resultado del inciso (a) sea válido en el caso general.

11. En un trabajo titulado *Population Waves*, Bernadelli [Ber] (véase también [Se]) postula la existencia hipotética de un escarabajo simplificado cuya vida natural es de 3 años. La hembra de esta especie tiene una tasa de supervivencia de $\frac{1}{2}$ en el primer año de vida, de $\frac{1}{3}$ del segundo al tercer año de vida y procrea un promedio de seis hembras antes de morir al final del tercer año. Podemos utilizar una matriz para demostrar la contribución que un escarabajo hembra hace, en sentido probabilístico, a la población femenina de la especie, al denotar con a_{ij} en la matriz $A = (a_{ij})$ a la contribución que un escarabajo hembra de edad j hará a la población femenina de edad i del siguiente año; es decir,

$$A = \begin{bmatrix} 0 & 0 & 6 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{bmatrix}.$$

a. La contribución que un escarabajo hembra hace a la población al cabo de 2 años se determina a partir de los elementos A^2, al cabo de 3 años a partir de A^3, y así sucesivamente. Construya A^2 y A^3 y trate de hacer un enunciado general sobre la contribución de un escarabajo hembra a la población en n años para cualquier valor entero positivo de n.

b. Con base en las conclusiones del inciso (a) describa lo que sucederá en años futuros a una población de estos escarabajos que inicialmente eran 6000 escarabajos hembra en cada uno de los tres grupos de edad.

c. Construya A^{-1} y describa su importancia para la población de esta especie.

12. El estudio de las cadenas alimentarias constituye un tema muy importante en la determinación de la propagación y acumulación de los contaminantes ambientales en la materia viva. Suponga que una cadena alimentaria consta de tres eslabones. El primero se compone de una vegetación de los tipos $v_1, v_2, \ldots, v_n$, que satisfacen todas las necesidades alimentarias de los herbívoros de las especies, $h_1, h_2, \ldots, h_m$, en el segundo eslabón. El tercer eslabón consta de carnívoros $c_1, c_2, \ldots, c_k$, cuyo abastecimiento depende por completo de los herbívoros del segundo eslabón. La coordenada a_{ij} de la matriz

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

representa el número total de plantas de tipo v_i ingeridas por los herbívoros de la especie h_j, mientras que b_{ij} en

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mk} \end{bmatrix}$$

describe el número de herbívoros de la especie h_i que son devorados por los animales de tipo c_j.

a. Demuestre que el número de plantas de tipo v_i que finalmente terminan en los animales de especie c_j está dado por el elemento del i-ésimo renglón y la j-ésima columna de la matriz AB.

b. ¿Qué importancia física tienen las matrices A^{-1}, B^{-1} y $(AB)^{-1} = B^{-1}A^{-1}$?

13. En la sección 3.5 descubrimos que la forma paramétrica $(x(t), y(t))$ de los polinomios cúbicos de Hermite $(x(0), y(0)) = (x_0, y_0)$ y $(x(1), y(1)) = (x_1, y_1)$ con los puntos guía $(x_0 + \alpha_0, y_0 + \beta_0)$ y $(x_1 + \alpha_1, y_1 + \beta_1)$, respectivamente, están dados por

$$x(t) = [2(x_0 - x_1) + (\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - \alpha_1 - 2\alpha_0]t^2 + \alpha_0 t + x_0$$

y

$$y(t) = [2(y_0 - y_1) + (\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - \beta_1 - 2\beta_0]t^2 + \beta_0 t + y_0.$$

Los polinomios cúbicos de Bézier tienen la forma

$$\hat{x}(t) = [2(x_0 - x_1) + 3(\alpha_0 + \alpha_1)]t^3 + [3(x_1 - x_0) - 3(\alpha_1 + 2\alpha_0)]t^2 + 3\alpha_0 t + x_0$$

y

$$\hat{y}(t) = [2(y_0 - y_1) + 3(\beta_0 + \beta_1)]t^3 + [3(y_1 - y_0) - 3(\beta_1 + 2\beta_0)]t^2 + 3\beta_0 t + y_0.$$

a. Demuestre que la matriz

$$A = \begin{bmatrix} 7 & 4 & 4 & 0 \\ -6 & -3 & -6 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

transforma los coeficientes del polinomio de Hermite en los del polinomio de Bézier.

b. Determine una matriz B que transforma los coeficientes del polinomio de Bézier en los del polinomio de Hermite.

14. Considere el sistema lineal de 2×2 $(A + iB)(\mathbf{x} + i\mathbf{y}) = \mathbf{c} + i\mathbf{d}$ con elementos complejos en forma de componentes:

$$(a_{11} + ib_{11})(x_1 + iy_1) + (a_{12} + ib_{12})(x_2 + iy_2) = c_1 + id_1,$$
$$(a_{21} + ib_{21})(x_1 + iy_1) + (a_{22} + ib_{22})(x_2 + iy_2) = c_2 + id_2.$$

a. Use las propiedades de los números complejos para convertir este sistema en el sistema lineal real equivalente de 4×4

$$A\mathbf{x} - B\mathbf{y} = \mathbf{c},$$
$$B\mathbf{x} + A\mathbf{y} = \mathbf{d}.$$

b. Resuelva el sistema lineal

$$(1 - 2i)(x_1 + iy_1) + (3 + 2i)(x_2 + iy_2) = 5 + 2i,$$
$$(2 + i)(x_1 + iy_1) + (4 + 3i)(x_2 + iy_2) = 4 - i.$$

6.4 Determinante de una matriz

El *determinante* de una matriz es un concepto fundamental del álgebra lineal con el cual se determina la existencia y la unicidad de los resultados de los sistemas de ecuaciones lineales. Para representar el determinante de una matriz A usaremos det A, aunque también se acostumbra utilizar la notación $|A|$.

Definición 6.14

a. Si $A = [a]$ es una matriz de 1×1, entonces det $A = a$.

b. Si A es una matriz de $n \times n$, el **menor** M_{ij} es el determinante de la submatriz $(n - 1) \times (n - 1)$ de A, que se obtiene al suprimir el i-ésimo renglón y la j-ésima columna de la matriz A.

c. El **cofactor** A_{ij} asociado a M_{ij} se define como $A_{ij} = (-1)^{i+j}M_{ij}$.

d. El **determinante** de la matriz A de $n \times n$, cuando $n > 1$, está dado por

$$\det A = \sum_{j=1}^{n} a_{ij} A_{ij} = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij}, \quad \text{para cada} \quad i = 1, 2, \ldots, n,$$

o bien por

$$\det A = \sum_{i=1}^{n} a_{ij} A_{ij} = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij}, \quad \text{para cada} \quad j = 1, 2, \ldots, n. \quad ■$$

Podemos demostrar (ver ejercicio 9) que, para calcular el determinante de una matriz general de $n \times n$ mediante esta definición, se requieren $O(n!)$ multiplicaciones/divisiones y sumas/restas. Incluso con valores de n relativamente pequeños, la cantidad de cálculos se torna inmanejable.

Al parecer hay $2n$ definiciones de det A, según el renglón o columnas que se escojan. Pero todas las definiciones arrojan el mismo resultado numérico. La flexibilidad de la definición se emplea en el siguiente ejemplo. Es más adecuado calcular det A a lo largo del renglón o en la columna que tengan más ceros.

EJEMPLO 1 Sea

$$A = \begin{bmatrix} 2 & -1 & 3 & 0 \\ 4 & -2 & 7 & 0 \\ -3 & -4 & 1 & 5 \\ 6 & -6 & 8 & 0 \end{bmatrix}.$$

Si queremos calcular det A, es más fácil efectuar la expansión alrededor de la cuarta columna:

$$\det A = a_{14}A_{14} + a_{24}A_{24} + a_{34}A_{34} + a_{44}A_{44} = 5A_{34} = -5M_{34}.$$

Al suprimir el tercer renglón y la cuarta columna, obtenemos

$$\det A = -5 \det \begin{bmatrix} 2 & -1 & 3 \\ 4 & -2 & 7 \\ 6 & -6 & 8 \end{bmatrix}$$

$$= -5 \left\{ 2 \det \begin{bmatrix} -2 & 7 \\ -6 & 8 \end{bmatrix} - (-1) \det \begin{bmatrix} 4 & 7 \\ 6 & 8 \end{bmatrix} + 3 \det \begin{bmatrix} 4 & -2 \\ 6 & -6 \end{bmatrix} \right\} = -30. \quad ■$$

En Maple, el determinante de una matriz se calcula por medio del comando `det(A);`.

Las siguientes propiedades sirven para relacionar los sistemas lineales y la eliminación gaussiana con los determinantes. La demostración de ellas se encuentra en cualquier libro de álgebra lineal (véase, por ejemplo, [ND, pp. 200-201].)

Teorema 6.15 Supóngase que A es una matriz de $n \times n$:

a. Si un renglón o columna cualquiera de A tiene exclusivamente elementos cero, entonces det $A = 0$.

b. Si A tiene dos renglones iguales o dos columnas iguales, entonces det $A = 0$.

c. Si obtenemos $\tilde{A}$ a partir de A por medio de la operación $(E_i) \leftrightarrow (E_j)$, con $i \neq j$, entonces det $\tilde{A} = -\det A$.

d. Si obtenemos $\tilde{A}$ a partir de A por medio de la operación $(\lambda E_i) \rightarrow (E_i)$, entonces det $\tilde{A} = \lambda \det A$.

e. Si obtenemos $\tilde{A}$ a partir de A por medio de la operación $(E_i + \lambda E_j) \rightarrow (E_i)$, con $i \neq j$, entonces det $\tilde{A} = \det A$.

f. Si B es también una matriz de $n \times n$, entonces det $AB = \det A \det B$.

g. det $A^t = \det A$.

h. Cuando A^{-1} existe, det $A^{-1} = (\det A)^{-1}$.

i. Si A es una matriz triangular superior, triangular inferior o diagonal, entonces det $A = \prod_{i=1}^{n} a_{ii}$. ■

Es facil calcular el determinante de una matriz en forma triangular, de modo que podemos simplificar el cálculo del determinante de cualquier matriz reduciendo primero la matriz a una forma triangular, y usando luego el inciso (I) del teorema para calcular el determinante de la matriz triangular.

EJEMPLO 2 Calcule el determinante de la matriz

$$A = \begin{bmatrix} 2 & 1 & -1 & 1 \\ 1 & 1 & 0 & 3 \\ -1 & 2 & 3 & -1 \\ 3 & -1 & -1 & 2 \end{bmatrix}$$

usando los incisos **(b)**, **(d)** y **(e)** del teorema 6.15, realizando los cálculos en Maple. La matriz A se define como

```
A:=matriz(4,4,[2,1,-1,1,1,1,0,3,-1,2,3,-1,3,-1,-1,2]);
```

La serie de operaciones en la tabla 6.2 produce la matriz

$$A8 = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & -0 & -13 \end{bmatrix}.$$

Por el inciso (I), $\det A8 = -39$, de modo que $\det A = 39$. ■

Tabla 6.2

Operación	Maple	Efecto
$\frac{1}{2}E_1 \to E_1$	`A1:= mulrow(A,1,0.5)`	$\det A1 = \frac{1}{2}\det A$
$E_2 - E_1 \to E_2$	`A2:= addrow(A1,1,2,-1)`	$\det A2 = \det A1 = \frac{1}{2}\det A$
$E_3 + E_1 \to E_3$	`A3:= addrow(A2,1,3,1)`	$\det A3 = \det A2 = \frac{1}{2}\det A$
$E_4 - 3E_1 \to E_4$	`A4:= addrow(A3,1,4,-3)`	$\det A4 = \det A3 = \frac{1}{2}\det A$
$2E_2 \to E_2$	`A5:= mulrow(A,2,2)`	$\det A5 = 2\det A4 = \det A$
$E_3 - \frac{5}{2}E_2 \to E_3$	`A6:= addrow(A,2,3,-2.5)`	$\det A6 = \det A5 = \det A$
$E_4 + \frac{5}{2}E_2 \to E_4$	`A7:= addrow(A,2,4,2.5)`	$\det A7 = \det A6 = \det A$
$E_3 \leftrightarrow E_4$	`A8:= swaprow(A,3,4)`	$\det A8 = -\det A7 = -\det A$

A continuación presentamos el resultado más importante que relaciona la no singularidad, la eliminación gaussiana, los sistemas lineales y los determinantes.

Teorema 6.16 Las afirmaciones que sigue son equivalentes para cualquier matriz A de $n \times n$:

a. La ecuación $A\mathbf{x} = \mathbf{0}$ tiene la solución única $\mathbf{x} = \mathbf{0}$.

b. El sistema $A\mathbf{x} = \mathbf{b}$ tiene la solución única para cualquier vector columna n-dimensional $\mathbf{b}$.

c. La matriz A es no singular; es decir, existe A^{-1}.

d. $\det A \neq 0$.

e. La eliminación gaussiana con intercambios de renglones puede efectuarse en el sistema $A\mathbf{x} = \mathbf{b}$ para cualquier vector $\mathbf{b}$ columna n-dimensional. ■

CONJUNTO DE EJERCICIOS 6.4

1. Use la definición 6.14 para calcular los determinantes de las siguientes matrices:

a. $\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & -1 \\ 3 & 1 & 1 \end{bmatrix}$ **b.** $\begin{bmatrix} 4 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & 2 & 3 \end{bmatrix}$

c. $\begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & 2 & -4 & -2 \\ 2 & 1 & 1 & 5 \\ -1 & 0 & -2 & -4 \end{bmatrix}$ **d.** $\begin{bmatrix} 2 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 2 & -1 & 3 & 1 \\ 3 & -1 & 4 & 3 \end{bmatrix}$

2. Repita el ejercicio 1 aplicando el método del ejemplo 2.

3. Calcule $\det A$, $\det B$, $\det AB$ y $\det BA$ para

$$A = \begin{bmatrix} 4 & 6 & 1 & -1 \\ 2 & 1 & 0 & \frac{1}{2} \\ 3 & 0 & 0 & 1 \\ 1 & -1 & 1 & 1 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

4. Sea A una matriz de 3×3. Demuestre que si $\tilde{A}$ es una matriz que se obtuvo de A por medio de una de las operaciones

$$(E_1) \leftrightarrow (E_2), \quad (E_1) \leftrightarrow (E_3) \quad \text{o} \quad (E_2) \leftrightarrow (E_3),$$

entonces $\det \tilde{A} = -\det A$.

5. Obtenga todos los valores de α que hacen que la siguiente matriz sea singular.

$$A = \begin{bmatrix} 1 & -1 & \alpha \\ 2 & 2 & 1 \\ 0 & \alpha & -\frac{3}{2} \end{bmatrix}.$$

6. Obtenga todos los valores de α que hacen que la siguiente matriz sea singular.

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 1 & \alpha & 1 \\ 2 & \alpha & -1 \end{bmatrix}$$

7. Obtenga todos los valores de α de modo que el siguiente sistema lineal no tenga solución.

$$\begin{aligned} 2x_1 - x_2 + 3x_3 &= 5, \\ 4x_1 + 2x_2 + 2x_3 &= 6, \\ -2x_1 + \alpha x_2 + 3x_3 &= 4. \end{aligned}$$

8. Obtenga todos los valores de α de manera que el siguiente sistema lineal tenga un número infinito de soluciones.

$$\begin{aligned} 2x_1 - x_2 + 3x_3 &= 5, \\ 4x_1 + 2x_2 + 2x_3 &= 6, \\ -2x_1 + \alpha x_2 + 3x_3 &= 1. \end{aligned}$$

9. Use inducción matemática para demostrar que, cuando $n > 1$, la evaluación del determinante de una matriz de $n \times n$ mediante la definición requiere $n! \sum_{k=1}^{n-1} \frac{1}{k!}$ multiplicaciones/divisiones y $n! - 1$ sumas/restas.

10. Demuestre que AB es no singular si y sólo si A y B son no singulares.

11. La solución mediante la **regla de Cramer** del sistema lineal

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3, \end{aligned}$$

tiene

$$x_1 = \frac{1}{D} \det \begin{bmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{bmatrix} \equiv \frac{D_1}{D},$$

$$x_2 = \frac{1}{D} \det \begin{bmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{bmatrix} \equiv \frac{D_2}{D},$$

y

$$x_3 = \frac{1}{D} \det \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{bmatrix} \equiv \frac{D_3}{D},$$

donde

$$D = \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

a. Obtenga la solución del sistema lineal

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &= 4, \\ x_1 - 2x_2 + x_3 &= 6, \\ x_1 - 12x_2 + 5x_3 &= 10, \end{aligned}$$

aplicando la regla de Cramer.

b. Demuestre que el sistema lineal

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &= 4, \\ x_1 - 2x_2 + x_3 &= 6, \\ -x_1 - 12x_2 + 5x_3 &= 9 \end{aligned}$$

no tiene solución. Calcule D_1, D_2 y D_3.

c. Demuestre que el sistema lineal

$$\begin{aligned} 2x_1 + 3x_2 - x_3 &= 4, \\ x_1 - 2x_2 + x_3 &= 6, \\ -x_1 - 12x_2 + 5x_3 &= 10 \end{aligned}$$

tiene un número infinito de soluciones. Calcule D_1, D_2 y D_3.

d. Demuestre que, si un sistema lineal de 3×3 con $D = 0$ tiene soluciones, entonces $D_1 = D_2 = D_3 = 0$.

e. Determine el número de multiplicaciones/divisiones y de sumas/restas que requiere la regla de Cramer en un sistema de 3×3.

12. a. Generalice la regla de Cramer a un sistema lineal de $n \times n$.

b. Use el resultado del ejercicio 9 para determinar el número de multiplicaciones/divisiones y de sumas/restas que requiere la regla de Cramer en un sistema de $n \times n$.

6.5 Factorización de matrices

La eliminación gaussiana es la principal herramienta en la solución directa de los sistemas de ecuaciones lineales; por ello, no debe sorprendernos que aparezca en otras formas. En la presente sección veremos que los pasos que se siguen para resolver un sistema de la forma $A\mathbf{x} = \mathbf{b}$, también pueden servir para factorizar una matriz en un producto matricial. La factorización es muy útil cuando presenta la forma $A = LU$, donde L es triangular inferior y U es triangular superior. No todas las matrices pueden factorizarse de este modo, pero es posible hacerlo con un gran número de las que se presentan con frecuencia en las aplicaciones.

En la sección 6.1 vimos que la eliminación gaussiana aplicada a un sistema lineal arbitrario $A\mathbf{x} = \mathbf{b}$ requiere $O(n^3/3)$ operaciones aritméticas para determinar $\mathbf{x}$. Si A ha sido factorizada en la forma triangular $A = LU$, entonces podemos encontrar $\mathbf{x}$ más fácilmente empleando un proceso de dos pasos. Primero usamos $\mathbf{y} = U\mathbf{x}$ y resolvemos el sistema $L\mathbf{y} = \mathbf{b}$ para $\mathbf{y}$. Puesto que L es triangular, para determinar $\mathbf{y}$ a partir de esta ecuación sólo necesitaremos $O(n^2)$ operaciones, o sea las mismas que en la parte de la eliminación gaussiana corresponde a la sustitución hacia atrás. Una vez que conocemos $\mathbf{y}$, el sistema triangular $U\mathbf{x} = \mathbf{y}$ requiere sólo $O(n^2)$ operaciones adicionales para encontrar la solución $\mathbf{x}$. Este hecho implica que la cantidad de operaciones necesarias para resolver el sistema $A\mathbf{x} = \mathbf{b}$ se reduzca de $O(2n^3/3)$ a $O(n^2)$. En los sistemas mayores de 100 por 100, ese resultado puede disminuir la cantidad de cálculos en más de 97%. No debe sorprendernos que las reducciones resultantes de la factorización matricial tengan un "costo" en cuanto al número de operaciones: para determinar las matrices específicas L y U se necesitan $O(n^3/3)$ operaciones. Pero una vez que se determina la factorización, podemos resolver en forma simplificada cualquier sistema que contenga la matriz A.

A fin de examinar cuáles matrices tienen una factorización LU y averiguar cómo se determina ésta, supondremos primero que la eliminación gaussiana puede efectuarse en el sistema $A\mathbf{x} = \mathbf{b}$ sin intercambio de renglones. Cuando empleamos la notación de la sección 6.1, lo anterior equivale a tener elementos pivote no cero $a_{ii}^{(i)}$ para toda $i = 1, 2, \ldots, n$.

El primer paso en el proceso de la eliminación gaussiana consiste en realizar, para toda $j = 2, 3, \ldots, n$, las operaciones.

$$(E_j - m_{j,1} E_1) \rightarrow (E_j), \quad \text{donde} \quad m_{j,1} = \frac{a_{j1}^{(1)}}{a_{11}^{(1)}}. \tag{6.8}$$

Estas operaciones transforman el sistema en otro en el cual todos los componentes de la primera columna situadas debajo de la diagonal son cero.

Podemos ver desde otro punto de vista el sistema de operaciones en (6.8). Esto lo logramos simultáneamente al multiplicar la matriz original A de la izquierda por la matriz

$$M^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -m_{21} & 1 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -m_{n1} & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Ésta recibe el nombre de **primera matriz gaussiana de transformación.** El producto de esta matriz lo denotamos con $A^{(1)} \equiv A$ por $A^{(2)}$ y con $\mathbf{b}$ por $\mathbf{b}^{(2)}$, así que

$$A^{(2)}\mathbf{x} = M^{(1)}A\mathbf{x} = M^{(1)}\mathbf{b} = \mathbf{b}^{(2)}.$$

De manera semejante construimos la matriz identidad $M^{(2)}$, en la que los elementos situados debajo de la diagonal en la segunda columna se reemplazan con los negativos de los multiplicadores

$$m_{j,2} = \frac{a_{j2}^{(2)}}{a_{22}^{(2)}}.$$

El producto de esta matriz con $A^{(2)}$ tiene ceros debajo de la diagonal de las dos primeras columnas y usamos

$$A^{(3)}\mathbf{x} = M^{(2)}A^{(2)}\mathbf{x} = M^{(2)}M^{(1)}A\mathbf{x} = M^{(2)}M^{(1)}\mathbf{b} = \mathbf{b}^{(3)}.$$

En general, con $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ ya formada, multiplicamos **por la k-ésima matriz de la transformación gaussiana**

$$M^{(k)} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & \ddots & \ddots & & & \vdots \\ \vdots & & \vdots & -m_{k+1,k} & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -m_{n,k} & 0 & \cdots & 0 & 1 \end{bmatrix}$$

para obtener

$$A^{(k+1)}\mathbf{x} = M^{(k)}A^{(k)}\mathbf{x} = M^{(k)}\cdots M^{(1)}A\mathbf{x} = M^{(k)}\mathbf{b}^{(k)} = \mathbf{b}^{(k+1)} = M^{(k)}\cdots M^{(1)}\mathbf{b}. \quad (6.9)$$

El proceso termina con la formación de $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$, donde $A^{(n)}$ es la matriz triangular superior.

$$A^{(n)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n}^{(n-1)} \\ 0 & \cdots & 0 & a_{nn}^{(n)} \end{bmatrix}$$

dada por

$$A^{(n)} = M^{(n-1)} M^{(n-2)} \cdots M^{(1)} A.$$

El proceso que hemos descrito forma la parte $U = A^{(n)}$ de la factorización matricial $A = LU$. Si queremos determinar la matriz triangular inferior L, primero debemos recordar la multiplicación de $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ mediante la transformación gaussiana $M^{(k)}$ con que obtuvimos (6.9):

$$A^{(k+1)}\mathbf{x} = M^{(k)}A^{(k)}\mathbf{x} = M^{(k)}\mathbf{b}^{(k)} = \mathbf{b}^{(k+1)},$$

donde $M^{(k)}$ genera las operaciones de renglón

$$(E_j - m_{j,k} E_k) \to (E_j), \quad \text{para} \quad j = k+1, \ldots, n.$$

Para revertir los efectos de esta transformación y volver a $A^{(k)}$ hay que realizar las operaciones $(E_j + m_{j,k} E_k) \to (E_j)$ para toda $j = k+1, \ldots, n$. Esto equivale a multiplicar por la inversa de la matriz $M^{(k)}$, la matriz

$$L^{(k)} = [M^{(k)}]^{-1} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & 0 & \ddots & \ddots & & \vdots \\ \vdots & & \vdots & m_{k+1,k} & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & m_{n,k} & 0 & \cdots 0 & 1 \end{bmatrix}.$$

En la factorización de A la matriz triangular inferior L es producto de las matrices $L^{(k)}$:

$$L = L^{(1)}L^{(2)} \cdots L^{(n-1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{bmatrix},$$

porque el producto de L con la matriz triangular superior $U = M^{(n-1)} \ldots M^{(2)} M^{(1)}A$ da

$$\begin{aligned} LU &= L^{(1)}L^{(2)} \cdots L^{(n-3)}L^{(n-2)}L^{(n-1)} \cdot M^{(n-1)}M^{(n-2)}M^{(n-3)} \cdots M^{(2)}M^{(1)}A \\ &= [M^{(1)}]^{-1}[M^{(2)}]^{-1} \cdots [M^{(n-2)}]^{-1}[M^{(n-1)}]^{-1} \cdot M^{(n-1)}M^{(n-2)} \cdots M^{(2)}M^{(1)}A = A. \end{aligned}$$

El teorema 6.17 se obtiene de las observaciones anteriores.

Teorema 6.17 Si podemos efectuar la eliminación gaussiana en el sistema lineal $A\mathbf{x} = \mathbf{b}$ sin intercambios de renglones, entonces podemos factorizar la matriz A en el producto de una matriz triangular inferior L y una matriz triangular superior U,

$$A = LU,$$

donde $m_{ji} = a_{ji}^{(i)}/a_{ii}^{(i)}$,

$$U = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n}^{(n-1)} \\ 0 & \cdots & 0 & a_{nn}^{(n)} \end{bmatrix}, \quad \text{y} \quad L = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{bmatrix}.$$ ■

EJEMPLO 1 En la sección 6.1 consideramos el sistema lineal

$$\begin{aligned} x_1 + x_2 \quad\quad + 3x_4 &= 4, \\ 2x_1 + x_2 - x_3 + x_4 &= 1, \\ 3x_1 - x_2 - x_3 + 2x_4 &= -3, \\ -x_1 + 2x_2 + 3x_3 - x_4 &= 4. \end{aligned}$$

La secuencia de operaciones $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, $(E_4 - (-1)E_1) \to (E_4)$, $(E_3 - 4E_2) \to (E_3)$, $(E_4 - (-3)E_2) \to (E_4)$ lo convierte en el sistema triangular

$$\begin{aligned} x_1 + x_2 \quad\quad + 3x_4 &= 4, \\ -x_2 - x_3 - 5x_4 &= -7, \\ 3x_3 + 13x_4 &= 13, \\ -13x_4 &= -13. \end{aligned}$$

Los multiplicadores m_{ij} y la matriz triangular superior producen la factorización

$$A = \begin{bmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} = LU.$$

Esta factorización nos permite fácilmente resolver el sistema que contiene la matriz A. Por ejemplo, para resolver

$$A\mathbf{x} = LU\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \\ 14 \\ -7 \end{bmatrix},$$

primero introducimos la sustitución $\mathbf{y} = U\mathbf{x}$. Luego $L\mathbf{y} = \mathbf{b}$, es decir,

$$LU\mathbf{x} = L\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \\ 14 \\ -7 \end{bmatrix}.$$

Este sistema se resuelve para $\mathbf{y}$ mediante un proceso de sustitución hacia adelante simple:

$$\begin{aligned} y_1 &= 8; & & \\ 2y_1 + y_2 &= 7, & \text{de donde} \quad & y_2 = 7 - 2y_1 = -9; \\ 3y_1 + 4y_2 + y_3 &= 14, & \text{de donde} \quad & y_3 = 14 - 3y_1 - 4y_2 = 26; \\ -y_1 - 3y_2 + y_4 &= -7, & \text{de donde} \quad & y_4 = -7 + y_1 + 3y_2 = -26. \end{aligned}$$

y entonces resolvemos $U\mathbf{x} = \mathbf{y}$ para $\mathbf{x}$, o sea la solución del sistema original; es decir,

$$\begin{bmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ -9 \\ 26 \\ -26 \end{bmatrix}.$$

Al emplear la sustitución hacia atrás, obtenemos $x_4 = 2$, $x_3 = 0$, $x_2 = -1$, $x_1 = 3$. ■

La factorización utilizada en el ejemplo 1 recibe el nombre de *método de Doolittle* y requiere que haya unos en la diagonal de L, lo que da como resultado la factorización descrita en el teorema 6.17. En la sección 6.6 veremos el *método de Crout*, factorización que requiere que los elementos diagonales de U sean uno, y el *método de Choleski*, que requiere que $l_{ii} = u_{ii}$ para toda i.

En el algoritmo 6.4 se incluye un procedimiento general con que se factorizan matrices y se convierten en un producto de matrices triangulares. Aunque se construyen las matrices nuevas L y U, podemos reemplazar los elementos correspondientes de A que ya no necesitemos con los valores generados.

El algoritmo 6.4 permite especificar la diagonal de L o de U.

ALGORITMO 6.4

Factorización *LU*

Para factorizar la matriz A de $n \times n = (a_{ij})$ en el producto de la matriz triangular inferior $L = (l_{ij})$ y en la matriz triangular superior $U = (u_{ij})$; es decir, $A = LU$, donde la diagonal principal de L o U consta de unos:

ENTRADA la dimensión n; los elementos a_{ij}, $1 \le i, j \le n$ de A, la diagonal $l_{11} = \cdots = l_{nn} = 1$ de L o la diagonal $u_{11} = \cdots = u_{nn} = 1$ de U.

SALIDA los elementos l_{ij}, $1 \le j \le i$, $1 \le i \le n$ de L y los elementos u_{ij}, $i \le j \le n$, $1 \le i \le n$ de U.

Paso 1 Seleccione l_{11} y u_{11} satisfaciendo $l_{11}u_{11} = a_{11}$.
Si $l_{11}u_{11} = 0$ entonces SALIDA ('factorización imposible');
PARAR.

Paso 2 Para $j = 2, \ldots, n$ tome $u_{1j} = a_{1j}/l_{11}$; *(primer renglón de U).*
$l_{j1} = a_{j1}/u_{11}$; *(primera columna de L).*

Paso 3 Para $i = 2, \ldots, n-1$ haga los pasos 4 y 5.

Paso 4 Seleccione l_{ii} y u_{ii} satisfaciendo $l_{ii}u_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik}u_{ki}$.
Si $l_{ii}u_{ii} = 0$ entonces SALIDA ('factorización imposible');
PARAR.

Paso 5 Para $j = i+1, \ldots, n$

$$\text{tome } u_{ij} = \frac{1}{l_{ii}}\left[a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right]; \quad (i\text{-ésimo renglón de } U).$$

$$l_{ji} = \frac{1}{u_{ii}}\left[a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki}\right]. \quad (i\text{-ésima columna de } L.)$$

Paso 6 Seleccione l_{nn} y u_{nn} tales que $l_{nn}u_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}u_{kn}$.
(*Nota: Si* $l_{nn}u_{nn} = 0$, *entonces* $A = LU$ *pero* A *es singular.*)

Paso 7 SALIDA (l_{ij} para $j = 1, \ldots, i$ e $i = 1, \ldots, n$);
SALIDA (u_{ij} para $j = i, \ldots, n$ e $i = 1, \ldots, n$);
PARAR. ■

Una vez terminada la factorización matricial, la solución de un sistema lineal de la forma $A\mathbf{x} = LU\mathbf{x} = \mathbf{b}$ se obtiene primero haciendo que $\mathbf{y} = U\mathbf{x}$ y luego determinando $\mathbf{y}$ en $L\mathbf{y} = \mathbf{b}$. Como L es triangular superior, tenemos

$$y_1 = \frac{b_1}{l_{11}},$$

y, para toda $i = 2, 3, \ldots, n$,

$$y_i = \frac{1}{l_{ii}}\left[b_i - \sum_{j=1}^{i-1} l_{ij}y_j\right].$$

Una vez que se tiene $\mathbf{y}$, resolvemos el sistema triangular $U\mathbf{x} = \mathbf{y}$ para $\mathbf{x}$ por medio de la sustitución hacia atrás usando las ecuaciones

$$x_n = \frac{y_n}{u_{nn}} \quad \text{y} \quad x_i = \frac{1}{u_{ii}}\left[y_i - \sum_{j=i+1}^{n} u_{ij}x_j\right].$$

En la explicación anterior supusimos que podemos resolver $A\mathbf{x} = \mathbf{b}$ por medio de la eliminación gaussiana sin realizar intercambios de renglones. En la práctica, esta factorización sólo es útil cuando no se requieren intercambios de renglones para controlar el error de redondeo que resulta de usar la aritmética de dígitos finitos. Por fortuna, a este tipo pertenecen muchos de los sistemas que encontramos al aplicar métodos de aproximación; pero es interesante señalar las modificaciones que deben hacerse cuando se requieren intercambios de renglones. Comenzaremos la exposición describiendo una clase de matrices que sirven para rearreglar, o permutar, los renglones de una matriz.

Una **matriz de permutación** P, de $n \times n$, se obtiene al rearreglar los renglones de la matriz identidad. Esto produce una matriz que contiene exactamente un elemento distinto de cero en cada renglón y en cada columna. Los elementos que no son cero son todos unos.

EJEMPLO 2 La matriz

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

es una matriz de permutación 3×3. Para cualquier matriz A de 3×3, al multiplicar el lado izquierdo por P se intercambian los renglones segundo y tercero de A:

$$PA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

De manera análoga, al multiplicar A del lado derecho por P se intercambian las columnas segunda y tercera de A. ■

La matrices de permutación tienen dos propiedades de gran utilidad que se relacionan con la eliminación gaussiana. La primera de ellas se explica en el ejemplo anterior y establece que, si $k_1, \ldots, k_n$ es una permutación de los enteros $1, \ldots, n$, y si la matriz de permutación $P = (p_{ij})$ está definida por

$$P_{ij} = \begin{cases} 1, & \text{si } j = k_i \\ 0, & \text{en otro caso.} \end{cases}$$

Entonces

(i) PA permuta los renglones de A; esto es,

$$PA = \begin{bmatrix} a_{k_1 1} & a_{k_1 2} & \cdots & a_{k_1 n} \\ a_{k_2 1} & a_{k_2 2} & \cdots & a_{k_2 n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k_n 1} & a_{k_n 2} & \cdots & a_{k_n n} \end{bmatrix}.$$

(ii) P^{-1} existe y $P^{-1} = P^t$.

Al final de la sección 6.4 vimos que, para cualquier matriz no singular A, podemos resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$ mediante la eliminación gaussiana, sin excluir la posibilidad de realizar intercambios de renglones. Si conociéramos los intercambios que se requieren para resolver el problema mediante la eliminación gaussiana, podríamos arreglar las ecuaciones originales en un orden que garantice que no se requieran intercambios de renglón. Por tanto, hay un rearreglo de las ecuaciones en el sistema que permite efectuar la eliminación gaussiana *sin* intercambios de renglones. Ello significa que, en una matriz no singular A, existe una matriz de permutación P para la cual podemos resolver el sistema

$$PA\mathbf{x} = P\mathbf{b}$$

sin hacer intercambios de renglones. Pero podemos factorizar esta matriz PA en

$$PA = LU,$$

donde L es triangular inferior y U es triangular superior. Dado que $P^{-1} = P^t$, tenemos la factorización

$$A = P^{-1}LU = (P^tL)U.$$

La matriz U aún es triangular superior, sin embargo, la matriz P^tL no es triangular inferior a menos que $P = I$.

EJEMPLO 3 Puesto que $a_{11} = 0$, la matriz

$$A = \begin{bmatrix} 0 & 1 & -1 & 1 \\ 1 & 1 & -1 & 2 \\ -1 & -1 & 1 & 0 \\ 1 & 2 & 0 & 2 \end{bmatrix}$$

no tiene una factorización LU. Pero si utilizamos el intercambio de renglones $(E_1) \leftrightarrow (E_2)$, seguido de $(E_3 + E_1) \rightarrow E_3$ y de $(E_4 - E_1) \rightarrow E_4$, obtendremos

$$\begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Entonces el intercambio de renglones $(E_3) \leftrightarrow (E_4)$, seguida de $(E_3 - E_2) \rightarrow E_3$, nos da la matriz

$$U = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

La matriz de permutación asociada al intercambio de renglones $(E_1) \leftrightarrow (E_2)$ seguida del intercambio de renglones $(E_3) \leftrightarrow (E_4)$ es

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

La eliminación gaussiana en PA se puede realizar sin intercambio de renglones para obtener la factorización LU de PA,

$$PA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix} = LU.$$

De modo que

$$A = P^{-1}LU = (P^tL)U = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix}. \qquad \blacksquare$$

Maple tiene el comando `LUdecomp` para calcular una factorización de la forma $A = PLU$ de la matriz A. Si ya ha sido creada la matriz, la llamada a la función

```
>U:=LUdecomp(A,P='G',L='H');
```

regresa la matriz triangular superior U como el valor de la función, además de regresar la matriz triangular L en H y la matriz de permutación P en G.

CONJUNTO DE EJERCICIOS 6.5

1. Resuelva los siguientes sistemas lineales:

a. $$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 3 & -1 \\ 0 & -2 & 1 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

b. $$\begin{bmatrix} 2 & 0 & 0 \\ -1 & 1 & 0 \\ 3 & 2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 0 \end{bmatrix}$$

2. Considere las siguientes matrices. Encuentre la matriz de permutación P, tal que PA se pueda factorizar en el producto LU, donde L es triangular inferior con unos en su diagonal y donde U es triangular superior para las siguientes matrices.

a. $A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & 0 \\ 0 & 1 & -1 \end{bmatrix}$

b. $A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & -2 & -1 \\ 1 & -1 & 1 \end{bmatrix}$

c. $A = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & 1 & 4 & 3 \\ 2 & -1 & 2 & 4 \\ 2 & -1 & 2 & 3 \end{bmatrix}$

d. $A = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & -1 \\ 1 & 2 & -1 & 3 \\ 1 & 1 & 2 & 0 \end{bmatrix}$

3. Factorice las siguientes matrices en la descomposición LU aplicando el algoritmo de factorización LU con $l_{ii} = 1$ para toda i.

a. $\begin{bmatrix} 2 & -1 & 1 \\ 3 & 3 & 9 \\ 3 & 3 & 5 \end{bmatrix}$

b. $\begin{bmatrix} 1.012 & -2.132 & 3.104 \\ -2.132 & 4.906 & -7.013 \\ 3.104 & -7.013 & 0.014 \end{bmatrix}$

c. $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 1.5 & 0 & 0 \\ 0 & -3 & 0.5 & 0 \\ 2 & -2 & 1 & 1 \end{bmatrix}$

d. $\begin{bmatrix} 2.1756 & 4.0231 & -2.1732 & 5.1967 \\ -4.0231 & 6.0000 & 0 & 1.1973 \\ -1.0000 & -5.2107 & 1.1111 & 0 \\ 6.0235 & 7.0000 & 0 & -4.1561 \end{bmatrix}$

4. Modifique el algoritmo de factorización LU de manera que sirva para resolver un sistema lineal; después, resuelva los siguientes sistemas lineales.

a. $2x_1 - x_2 + x_3 = -1,$
$3x_1 + 3x_2 + 9x_3 = 0,$
$3x_1 + 3x_2 + 5x_3 = 4.$

b. $1.012x_1 - 2.132x_2 + 3.104x_3 = 1.984,$
$-2.132x_1 + 4.096x_2 - 7.013x_3 = -5.049,$
$3.104x_1 - 7.013x_2 + 0.014x_3 = -3.895.$

c. $2x_1 = 3,$
$x_1 + 1.5x_2 = 4.5,$
$- 3x_2 + 0.5x_3 = -6.6,$
$2x_1 - 2x_2 + x_3 + x_4 = 0.8,$

d. $2.1756x_1 + 4.0231x_2 - 2.1732x_3 + 5.1967x_4 = 17.102,$
$-4.0231x_1 + 6.0000x_2 + 1.1973x_4 = -6.1593,$
$-1.0000x_1 - 5.2107x_2 + 1.1111x_3 = 3.0004.$
$6.0235x_1 + 7.0000x_2 - 4.1561x_4 = 0.0000.$

5. Obtenga factorizaciones de la forma $A = P^tLU$ para las siguientes matrices.

a. $A = \begin{bmatrix} 0 & 2 & 3 \\ 1 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}$ **b.** $A = \begin{bmatrix} 1 & 2 & -1 \\ 1 & 2 & 3 \\ 2 & -1 & 4 \end{bmatrix}$

c. $A = \begin{bmatrix} 1 & -2 & 3 & 0 \\ 3 & -6 & 9 & 3 \\ 2 & 1 & 4 & 1 \\ 1 & -2 & 2 & -2 \end{bmatrix}$ **d.** $A = \begin{bmatrix} 1 & -2 & 3 & 0 \\ 1 & -2 & 3 & 1 \\ 1 & -2 & 2 & -2 \\ 2 & 1 & 3 & -1 \end{bmatrix}$

6. Suponga que $A = P^tLU$ donde P es la matriz de permutación, L es una matriz triangular inferior con unos en la diagonal y U es una matriz triangular superior.

a. Cuente el número de operaciones necesarias para calcular P^tLU para una matriz A dada.

b. Demuestre que, si P contiene k intercambios de renglones, entonces

$$\det P = \det P^t = (-1)^k.$$

c. Use $\det A = \det P^t \det L \det U = (-1)^k \det U$ para contar el número de opciones con que se determina $\det A$ mediante la factorización.

d. Calcule $\det A$ y cuente el número de operaciones cuando

$$A = \begin{bmatrix} 0 & 2 & 1 & 4 & -1 & 3 \\ 1 & 2 & -1 & 3 & 4 & 0 \\ 0 & 1 & 1 & -1 & 2 & -1 \\ 2 & 3 & -4 & 2 & 0 & 5 \\ 1 & 1 & 1 & 3 & 0 & 2 \\ -1 & -1 & 2 & -1 & 2 & 0 \end{bmatrix}.$$

7. a. Demuestre que el algoritmo de factorización LU requiere $\frac{1}{3}n^3 - \frac{1}{3}n$ multiplicaciones/divisiones y $\frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n$ sumas/restas.

b. Demuestre que para resolver $L\mathbf{y} = \mathbf{b}$, donde L es una matriz triangular inferior con $l_{ii} = 1$ para toda i, se requieren $\frac{1}{2}n^2 - \frac{1}{2}n$ multiplicaciones/divisiones y $\frac{1}{2}n^2 - \frac{1}{2}n$ sumas/restas.

c. Demuestre que para resolver $A\mathbf{x} = \mathbf{b}$ factorizando primero A en $A = LU$ y luego resolviendo $L\mathbf{y} = \mathbf{b}$ y $U\mathbf{x} = \mathbf{y}$ se requiere la misma cantidad de operaciones que las del algoritmo de la eliminación gaussiana 6.1.

d. Cuente el número de operaciones necesarias para resolver m sistemas lineales $A\mathbf{x}^{(k)} = \mathbf{b}^{(k)}$ para $k = 1, \ldots, m$ factorizando primero A y aplicando después el método del inciso (c) m veces.

6.6 Tipos especiales de matrices

En esta sección nos ocuparemos de dos clases de matrices en las cuales podemos practicar la eliminación gaussiana sin intercambios de renglones. La primera clase se describe en la siguiente definición.

Definición 6.18 Se dice que la matriz A de $n \times n$ es **estrictamente diagonal dominante** cuando

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

es aplicable a toda $i = 1, 2, \ldots, n$. ■

EJEMPLO 1 Consideremos las matrices

$$A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{bmatrix}.$$

La matriz no simétrica A es estrictamente diagonal dominante porque $|7| > |2| + |0|$, $|5| > |3| + |-1|$ y $|-6| > |0| + |5|$. La matriz simétrica B no es estrictamente diagonal dominante ya que, por ejemplo, en el tercer renglón el valor absoluto del elemento diagonal es $|6| < |4| + |-3| = 7$. Es interesante señalar que A^t no es estrictamente diagonal dominante y, por supuesto, tampoco lo es $B^t = B$. ■

El siguiente teorema se utilizó en la sección 3.4 a fin de asegurarnos de que los sistemas lineales necesarios para determinar los interpolantes de trazadores cúbicos tuvieran soluciones únicas.

Teorema 6.19 Una matriz A estrictamente diagonal dominante es no singular. Más aún, en este caso podemos realizar la eliminación gaussiana de cualquier sistema lineal de la forma $A\mathbf{x} = \mathbf{b}$ para obtener su solución única sin intercambios de renglones ni columnas y los cálculos son estables respecto al crecimiento de los errores de redondeo. ■

Demostración Primero aplicamos la prueba por contradicción para demostrar que A es no singular. Consideremos el sistema lineal descrito por $A\mathbf{x} = \mathbf{0}$ y supongamos que este sistema tiene una solución $\mathbf{x} = (x_i)$ no cero. Sea k un índice para el cual

$$0 < |x_k| = \max_{1 \leq j \leq n} |x_j|.$$

Dado que $\sum_{j=1}^{n} a_{ij}x_j = 0$ para toda $i = 1, 2, \ldots, n$, tendremos, cuando $i = k$,

$$a_{kk}x_k = -\sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj}x_j.$$

Esto implica que

$$|a_{kk}||x_k| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}||x_j|,$$

o bien

$$|a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|.$$

Esta desigualdad contradice el dominio diagonal estricto de A. En consecuencia, la única solución de $A\mathbf{x} = \mathbf{0}$, es $\mathbf{x} = \mathbf{0}$, condición que según se prueba en el teorema 6.16 equivale a la no singularidad de A.

Para demostrar que podemos efectuar la eliminación gaussiana sin intercambios de renglones, probaremos que cada una de las matrices $A^{(2)}, A^{(3)}, \ldots, A^{(n)}$ que se generan con ese proceso (y que se describen en la sección 6.5) tiene dominio diagonal estricto.

Puesto que A es estrictamente diagonal dominante, podemos formar $a_{11} \neq 0$ y $A^{(2)}$. Para toda $i = 2, 3, \ldots, n$,

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{1j}^{(1)}a_{i1}^{(1)}}{a_{11}^{(1)}}, \qquad \text{para} \qquad 2 \leq j \leq n.$$

Dado que $a_{i1}^{(2)} = 0$,

$$\begin{aligned}
\sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}^{(2)}| &= \sum_{\substack{j=2 \\ j\neq i}}^{n} \left| a_{ij}^{(1)} - \frac{a_{1j}^{(1)}a_{i1}^{(1)}}{a_{11}^{(1)}} \right| \\
&\leq \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{ij}^{(1)}| + \sum_{\substack{j=2 \\ j\neq i}}^{n} \left| \frac{a_{1j}^{(1)}a_{i1}^{(1)}}{a_{11}^{(1)}} \right| \\
&< |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|} \sum_{\substack{j=2 \\ j\neq i}}^{n} |a_{1j}^{(1)}| \\
&< |a_{ii}^{(1)}| - |a_{i1}^{(1)}| + \frac{|a_{i1}^{(1)}|}{|a_{11}^{(1)}|} (|a_{11}^{(1)}| - |a_{1i}^{(1)}|) \\
&= |a_{ii}^{(1)}| - \frac{|a_{i1}^{(1)}||a_{1i}^{(1)}|}{|a_{11}^{(1)}|} \\
&\leq \left| a_{ii}^{(1)} - \frac{a_{i1}^{(1)}a_{1i}^{(1)}}{|a_{11}^{(1)}|} \right| = |a_{ii}^{(2)}|.
\end{aligned}$$

Por tanto, el dominio diagonal estricto se establece en los renglones $2, \ldots, n$. Y como el primer renglón de $A^{(2)}$ y de A son iguales, $A^{(2)}$ será estrictamente diagonal dominante.

Este proceso continúa hasta obtener una $A^{(n)}$ estrictamente diagonal dominante. Ello significa que todos los elementos diagonales son no cero y, por tanto, podemos realizar la eliminación gaussiana sin intercambios de renglones.

La demostración de las estabilidad de este procedimiento viene en [We]. ■ ■ ■

La siguiente clase especial de matrices recibe el nombre de *definidas positivas*.

Definición 6.20 Una matriz A es **definida positiva** si es simétrica y si $\mathbf{x}^t A\mathbf{x} > 0$ para todo vector columna n dimensional $\mathbf{x} \neq \mathbf{0}$. ■

No todos los autores exigen simetría en una matriz definida positiva. Por ejemplo, Golub y Van Loan [GV], cuyo libro es clásico en los métodos matriciales, requieren sólo que $\mathbf{x}^t A\mathbf{x} > 0$ para cada $\mathbf{x} \neq \mathbf{0}$. Las matrices que llamamos definidas positivas reciben el nombre de definidas positivas simétricas en [GV]. Hay que tener presente esta diferencia de nomenclatura cuando utilice material de otras fuentes.

Si queremos ser muy precisos, la definición 6.20 debería especificar que la matriz de 1×1 generada con la operación $\mathbf{x}^t A\mathbf{x}$ tiene un valor positivo en su único elemento, ya que la operación se realiza así:

$$\mathbf{x}^t A\mathbf{x} = [x_1, x_2, \ldots, x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= [x_1, x_2, \ldots, x_n] \begin{bmatrix} \sum_{j=1}^n a_{1j} x_j \\ \sum_{j=1}^n a_{2j} x_j \\ \vdots \\ \sum_{j=1}^n a_{nj} x_j \end{bmatrix} = \left[\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \right].$$

EJEMPLO 2 La matriz

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Es definida positiva; suponiendo que $\mathbf{x}$ sea un vector columna de tres dimensiones, entonces

$$\mathbf{x}^t A\mathbf{x} = [x_1, x_2, x_3] \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= [x_1, x_2, x_3] \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{bmatrix}$$

$$= 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2.$$

Al rearreglar los términos obtenemos

$$\begin{aligned} \mathbf{x}^t A\mathbf{x} &= x_1^2 + (x_1^2 - 2x_1x_2 + x_2^2) + (x_2^2 - 2x_1x_3 + x_3^2) + x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 \end{aligned}$$

y

$$x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 > 0,$$

a menos que $x_1 = x_2 = x_3 = 0$. ■

Del ejemplo 2 podemos ver de manera evidente la dificultad de usar la definición para determinar si una matriz es definida positiva. Por fortuna contamos con criterios que son más fáciles de verificar y que se explican en el capítulo 9 para identificar los miembros de esta clase tan importante de matrices. El siguiente resultado contiene algunas de las condiciones que sirven para prescindir de ciertas matrices.

Teorema 6.21 Si A es una matriz definida positiva de $n \times n$, entonces

a. A es no singular;

b. $a_{ii} > 0$, para cada $i = 1, 2, \ldots, n$;

c. $\max_{1\le k,j\le n} |a_{kj}| \le \max_{1\le i\le n} |a_{ii}|$;

d. $(a_{ij})^2 < a_{ii}a_{jj}$ para cada $i \neq j$. ■

Demostración

a. Si $\mathbf{x}$ satisface $A\mathbf{x} = \mathbf{0}$, entonces $\mathbf{x}^t A\mathbf{x} = \mathbf{0}$. Esto contradice la suposición de que A es definida positiva. En consecuencia, $A\mathbf{x} = \mathbf{0}$ tiene sólo la solución cero y A es no singular.

b. Para una i cualquiera, definamos $\mathbf{x} = (x_j)$ por $x_i = 1$ y por $x_j = 0$, si $j \neq i$. Dado que $\mathbf{x} \neq \mathbf{0}$,

$$0 < \mathbf{x}^t A\mathbf{x} = a_{ii}.$$

c. Para $k \neq j$, definimos $\mathbf{x} = (x_i)$ por medio de

$$x_i = \begin{cases} 0, & \text{si } i \neq j \text{ e } i \neq k, \\ 1, & \text{si } i = j, \\ -1, & \text{si } i = k. \end{cases}$$

Puesto que $\mathbf{x} \neq \mathbf{0}$,

$$0 < \mathbf{x}^t A\mathbf{x} = a_{jj} + a_{kk} - a_{jk} - a_{kj}.$$

Pero $A^t = A$, así que $a_{jk} = a_{kj}$ y

$$2a_{kj} < a_{jj} + a_{kk}. \tag{6.10}$$

A continuación definimos $\mathbf{z} = (z_i)$ por medio de

$$z_i = \begin{cases} 0, & \text{si } i \neq j \text{ e } i \neq k, \\ 1, & \text{si } i = j \text{ o } i = k. \end{cases}$$

Entonces $\mathbf{z}^t A\mathbf{z} > 0$, así que

$$-2a_{kj} < a_{kk} + a_{jj}. \qquad (6.11)$$

Las ecuaciones (6.10) y (6.11) implican que para toda $k \neq j$,

$$|a_{kj}| < \frac{a_{kk} + a_{jj}}{2} \leq \max_{1\leq i\leq n} |a_{ii}|, \qquad \text{entonces} \qquad \max_{1\leq k,j\leq n} |a_{kj}| \leq \max_{1\leq i\leq n} |a_{ii}|.$$

d. Para $i \neq j$, definimos $\mathbf{x} = (x_k)$ por medio de

$$x_k = \begin{cases} 0, & \text{si } k \neq j \text{ y } k \neq i, \\ \alpha, & \text{si } k = i, \\ 1, & \text{si } k = j. \end{cases}$$

Donde α representa un número real arbitrario. Dado que $\mathbf{x} \neq \mathbf{0}$,

$$0 < \mathbf{x}^t A\mathbf{x} = a_{ii}\alpha^2 + 2a_{ij}\alpha + a_{jj}.$$

Como polinomio cuadrático en α sin raíces reales, el discriminante de $P(\alpha) = a_{ii}\alpha^2 + 2a_{ij}\alpha + a_{jj}$ debe ser negativo. Por tanto,

$$4a_{ij}^2 - 4a_{ii}a_{jj} < 0 \quad \text{y} \quad a_{ij}^2 < a_{ii}a_{jj}.$$ ■ ■ ■

Aunque el teorema 6.21 da algunas condiciones importantes que deben aplicarse a las matrices definidas positivas, no garantiza que una matriz que las cumpla sea de este tipo. El siguiente concepto servirá para establecer una condición necesaria y suficiente.

Definición 6.22 Una **primera submatriz principal** de una matriz A es la que tiene la forma

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix},$$

para alguna $1 \leq k \leq n$. ■

Una demostración del siguiente resultado se encuentra en [Stew1, p. 250].

Teorema 6.23 Una matriz simétrica A es definida positiva si y sólo si sus primeras submatrices principales tienen determinante positivo. ■

EJEMPLO 3 En el ejemplo 2 usamos la definición para demostrar que la matriz simétrica

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & -2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

es definida positiva. Para confirmar esto usando el teorema 6.23 nótese que

$$\det A_1 = \det [2] = 2 > 0,$$
$$\det A_2 = \det \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = 4 - 1 = 3 > 0,$$

y

$$\det A_3 = \det \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} = 2 \det \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} - (-1) \det \begin{bmatrix} -1 & -1 \\ 0 & 2 \end{bmatrix}$$
$$= 2(4-1) + (-2+0) = 4 > 0.$$ ■

El comando de Maple

```
>definite(A,positive_def);
```

devuelve un *verdadero* o *falso* como indicación de que A sea definida positiva o no. Para obtener un resultado *verdadero* se requiere la simetría.

El siguiente resultado amplía el inciso (a) del teorema 6.21 y corresponde a los resultados estrictamente dominantes en sentido diagonal que se incluyen en el teorema 6.19. No ofreceremos una demostración de este teorema, pues hay que introducir una terminología y resultados que no se necesitan en ninguna otra aplicación. El desarrollo y la demostración vienen en [We, p. 120 ss.].

Teorema 6.24 La matriz simétrica A es definida positiva si y sólo si la eliminación gaussiana sin intercambios de renglones puede efectuarse en el sistema lineal $A\mathbf{x} = \mathbf{b}$ con todos los elementos pivote positivos. Además, en este caso los cálculos son estables respecto al crecimiento de los errores de redondeo. ■

En los siguientes corolarios se incluyen algunos hechos interesantes que se descruben al construir la demostración del teorema 6.24.

Corolario 6.25 La matriz A es definida positiva si y sólo si A puede factorizares en la forma LDL^t, donde L es una matriz triangular inferior con unos en su diagonal y D es una matriz diagonal con elementos positivos a lo largo de la diagonal. ■

Corolario 6.26 La matriz A es definida positiva si y sólo si A puede factorizarse en la forma LL^t, donde L es una matriz triangular inferior con elementos distintos de cero en su diagonal. ■

La matriz L de corolario 6.26 no es la misma que la del corolario 6.25. En el ejercicio 26 se presenta una relación entre ellas.

El algoritmo 6.5 se basa en el algoritmo 6.4 de factorización LU y ofrece la factorización LDL^t que describimos en el corolario 6.25.

ALGORITMO 6.5

Factorización LDL^t

Para factorizar la matriz definida positiva A de $n \times n$, en la forma LDL^t, donde L es una matriz triangular inferior con unos a lo largo de la diagonal y donde D es una matriz diagonal con elementos positivos en la diagonal:

ENTRADA la dimensión n; elementos a_{ij}, para $1 \leq i, j \leq n$ de A.

SALIDA los elementos l_{ij}, para $1 \leq j < i$ y $1 \leq i \leq n$ de L y d_i, para $1 \leq i \leq n$ de D.

Paso 1 Para $i = 1, \ldots, n$ haga los pasos 2-4.

Paso 2 Para $j = 1, \ldots, i-1$, tome $v_j = l_{ij}d_j$.

Paso 3 Tome $d_i = a_{ii} - \sum_{j=1}^{i-1} l_{ij}v_j$.

Paso 4 Para $j = i+1, \ldots, n$ tome $l_{ji} = (a_{ji} - \sum_{k=1}^{i-1} l_{jk}v_k)/d_i$.

Paso 5 SALIDA (l_{ij} para $j = 1, \ldots, i-1$ e $i = 1, \ldots, n$);
SALIDA (d_i para $i = 1, \ldots, n$);
PARAR. ■

El corolario 6.25 tiene un equivalente cuando A es simétrica, pero no necesariamente definida positiva. Este resultado se presta a muchas aplicaciones, porque las matrices son comunes y fáciles de reconocer.

Corolario 6.27 Sea A una matriz simétrica de $n \times n$ a la cual puede aplicarse la eliminación gaussiana sin intercambios de renglones. Entonces, A puede factorizarse en LDL^t, donde L es una matriz triangular inferior con unos en su diagonal y donde D es una matriz diagonal con $a_{11}^{(1)}, \ldots, a_{nn}^{(n)}$ en su diagonal. ■

El algoritmo 6.5 se puede modificar fácilmente para factorizar las matrices simétricas descritas en el corolario 6.27. Basta hacer una inspección para asegurarse de que los elementos diagonales son distintos de cero. El algoritmo 6.6 de Choleski produce la factorización LL^t que se describe en el corolario 6.26.

ALGORITMO 6.6

Algoritmo de Choleski

Para factorizar la matriz definida positiva A de $n \times n$ en LL^t, donde L es una matriz triangular inferior:

ENTRADA la dimensión n; los elementos a_{ij}, para $1 \leq i, j \leq n$ de A.

SALIDA los elementos l_{ij}, para $1 \leq j \leq i$ y para $1 \leq i \leq n$ de L. (*Los elementos de $U = L^t$ son $u_{ij} = l_{ji}$, para $i \leq j \leq n$ y para $1 \leq i \leq n$.*)

Paso 1 Tome $l_{11} = \sqrt{a_{11}}$.

Paso 2 Para $j = 2, \ldots, n$, tome $l_{j1} = a_{j1}/l_{11}$.

Paso 3 Para $i = 2, \ldots, n - 1$ haga los pasos 4 y 5.

Paso 4 Tome $l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\right)^{1/2}$.

Paso 5 Para $j = i + 1, \ldots, n$

Tome $l_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}\, l_{ik}\right)/l_{ii}$.

Paso 6 Tome $l_{nn} = \left(a_{nn} - \sum_{k=1}^{n-1} l_{nk}^2\right)^{1/2}$.

Paso 7 SALIDA (l_{ij} para $j = 1, \ldots, i$ e $i = 1, \ldots, n$);
PARAR. ■

Maple calcula la factorización de Choleski de A con el enunciado;

```
>L:=cholesky(A);
```

EJEMPLO 4 La matriz

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{bmatrix}$$

es definida positiva. La factorización LDL^t de A dada en el algoritmo 6.5 es

$$A = LDL^t = \begin{bmatrix} 1 & 0 & 0 \\ -0.25 & 1 & 0 \\ 0.25 & 0.75 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -0.25 & 0.25 \\ 0 & 1 & 0.75 \\ 0 & 0 & 1 \end{bmatrix},$$

y el algoritmo 6.6 de Choleski produce la factorización

$$A = LL^t = \begin{bmatrix} 2 & 0 & 0 \\ -0.5 & 2 & 0 \\ 0.5 & 1.5 & 1 \end{bmatrix} \begin{bmatrix} 2 & -0.5 & 0.5 \\ 0 & 2 & 1.5 \\ 0 & 0 & 1 \end{bmatrix}.$$ ■

La factorización LDL^t que se describe en el algoritmo 6.5 requiere $n^3/6 + n^2 - 7n/6$ multiplicaciones/divisiones y $n^3/6 - n/6$ sumas/restas. La factorización LL^t de una matriz definida positiva de Choleski requiere apenas $n^3/6 + n^2/2 - 2n/3$ multiplicaciones/divisiones y $n^3/6 - n/6$ sumas/restas. No obstante, la ventaja computacional de la factorización de Choleski es engañosa, porque hay que extraer n raíces cuadradas. Pero la cantidad de operaciones necesarias para calcularla es un factor lineal de n y su importancia disminuirá al aumentar n.

El algoritmo 6.5 ofrece un método estable para factorizar una matriz definida positiva en la forma $A = LDL^t$, pero hay que modificarlo para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$. Si queremos hacer esto, suprimimos la proposición PARAR en el paso 5 del algoritmo y agregamos los siguientes pasos para resolver el sistema triangular inferior $L\mathbf{y} = \mathbf{b}$:

Paso 6 Tome $y_1 = b_1$.

Paso 7 Para $i = 2, \ldots, n$ tome $y_i = b_i - \sum_{j=1}^{i-1} l_{ij} y_j$.

Entonces podemos resolver el sistema lineal $D\mathbf{z} = \mathbf{y}$ por medio de

Paso 8 Para $i = 1, \ldots, n$ tome $z_i = y_i/d_i$.

Finalmente, el sistema triangular superior $L^t\mathbf{x} = \mathbf{z}$ se resuelve a través de los pasos dados por

Paso 9 Tome $x_n = z_n$.

Paso 10 Para $i = n - 1, \ldots, 1$ tome $x_i = z_i - \sum_{j=i+1}^{n} l_{ji} x_j$.

Paso 11 SALIDA (x_i para $i = 1, \ldots, n$);
PARAR.

En la tabla 6.3 se incluyen las operaciones adicionales necesarias para resolver el sistema lineal.

Tabla 6.3

Paso	Multiplicaciones/Divisiones	Sumas/Restas
6	0	0
7	$n(n-1)/2$	$n(n-1)/2$
8	n	0
9	0	0
10	$n(n-1)/2$	$n(n-1)/2$
Total	n^2	$n^2 - n$

Si se prefiere la factorización de Choleski dada en el algoritmo 6.6 para resolver el sistema $A\mathbf{x} = \mathbf{b}$ se utilizan los siguientes pasos adicionales. Primero se suprime la proposición PARAR en el paso 7. Después se agrega

Paso 8 tome $y_1 = b_1/l_{11}$.

Paso 9 Para $i = 2, \ldots, n$ tome $y_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij} y_j\right)/l_{ii}$.

Paso 10 tome $x_n = y_n/l_{nn}$.

Paso 11 Para $i = n - 1, \ldots, 1$ tome $x_i = \left(y_i - \sum_{j=i+1}^{n} l_{ji} x_j\right)/l_{ii}$.

Paso 12 SALIDA (x_i para $i = 1, \ldots, n$);
PARAR.

Los pasos 8-12 requieren $n^2 + n$ multiplicaciones/divisiones y $n^2 - n$ sumas/restas.

La última clase de matrices consideradas se denominan *matrices de banda*. En la generalidad de las aplicaciones estas matrices son estrictamente dominantes en sentido diagonal o definidas positivas.

Definición 6.28 Una matriz de $n \times n$ recibe el nombre de **matriz de banda** si existen los enteros p y q con $1 < p, q < n$, que tienen la propiedad de que $a_{ij} = 0$ siempre que $i + p \leq j$ o $j + q \leq i$. El **ancho de banda** de este tipo de matrices se define como $w = p + q - 1$. ■

Por ejemplo, la matriz

$$A = \begin{bmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & -5 & -6 \end{bmatrix}$$

es una matriz de banda con $p = q = 2$ y con ancho de banda $2 + 2 - 1 = 3$.

La definición de la matriz de banda hace que estas matrices concentren todos sus elementos distintos de cero alrededor de la diagonal. Dos casos especiales de matrices de banda que ocurren a menudo en la práctica tienen $p = q = 2$ y $p = q = 4$.

Las matrices de ancho de banda 3, que se presentan cuando $p = q = 2$, se les llama **tridiagonales** por tener la forma

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & & \vdots \\ 0 & a_{32} & a_{33} & a_{34} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

En el capítulo 11 se examinan las matrices tridiagonales al tratar del estudio de las aproximaciones lineales fragmentarias a los problemas de valor de frontera. Utilizaremos el caso $p = q = 4$ en la solución de esos problemas cuando las funciones de aproximación adoptan la forma de trazadores cúbicos.

Los algoritmos de factorización pueden simplificarse considerablemente en el caso de las matrices de banda, porque una gran cantidad de ceros aparecen en ellas en patrones regulares. Es muy interesante señalar la forma que en este caso asume el método de Crout o el de Doolittle.

Para ilustrar lo anterior, supóngase que podemos factorizar una matriz tridiagonal A en las matrices triangulares L y U. Puesto que A tiene sólo $(3n - 2)$ elementos distintos de cero, habrá apenas $(3n - 2)$ condiciones aplicables para determinar los elementos de L y U, naturalmente a condición de que también se obtengan los elementos cero de A. Supóngase que podemos encontrar las matrices en la forma

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & l_{n,n-1} & l_{nn} \end{bmatrix} \quad \text{y} \quad U = \begin{bmatrix} 1 & u_{12} & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & u_{n-1,n} \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}.$$

Hay $(2n - 1)$ elementos indeterminados de L y $(n - 1)$ elementos indeterminados de U, que suman el número total de condiciones, $(3n - 2)$. Los elementos cero de A se obtienen automáticamente.

La multiplicación que incluye $A = LU$ nos da, además de los elementos cero,

$$a_{11} = l_{11};$$

$$a_{i,i-1} = l_{i,i-1}, \quad \text{para cada } i = 2, 3, \ldots, n; \tag{6.12}$$

$$a_{ii} = l_{i,i-1}u_{i-1,i} + l_{ii}, \quad \text{para cada } i = 2, 3, \ldots, n; \tag{6.13}$$

y

$$a_{i,i+1} = l_{ii}u_{i,i+1}, \qquad \text{para cada} \qquad i = 1, 2, \ldots, n-1. \tag{6.14}$$

Una solución de este sistema se obtiene aplicando primero la ecuación (6.12) para obtener el término fuera de la diagonal de L y luego las ecuaciones (6.13) y (6.14) para obtener alternativamente el resto de los elementos de U y de L. Estos pueden guardarse en los elementos correspondientes de A.

El algoritmo 6.7 resuelve un sistema de ecuaciones lineales de $n \times n$ cuya matriz de coeficientes es tridiagonal. Sólo requiere $(5n - 4)$ multiplicaciones/divisiones y $(3n - 3)$ sumas/restas. En consecuencia, ofrece una importante ventaja computacional sobre los métodos que no toman en cuenta la tridiagonalidad de la matriz.

ALGORITMO 6.7

Factorización de Crout de sistemas lineales tridiagonales

Para resolver el sistema lineal de $n \times n$

$$\begin{array}{llll}
E_1: & a_{11}x_1 + a_{12}x_2 & & = a_{1,n+1}, \\
E_2: & a_{21}x_1 + a_{22}x_2 + a_{23}x_3 & & = a_{2,n+1}, \\
\vdots & \qquad\qquad \vdots & & \quad \vdots \\
E_{n-1}: & a_{n-1,n-2}x_{n-2} + a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n & & = a_{n-1,n+1}, \\
E_n: & \qquad a_{n,n-1}x_{n-1} + a_{nn}x_n & & = a_{n,n+1},
\end{array}$$

que se supone tiene una solución única:

ENTRADA la dimensión n; los elementos de A.

SALIDA la solución $x_1, \ldots, x_n$.

(*Use los pasos 1-3 y resuelva* $L\mathbf{z} = \mathbf{b}$.)

Paso 1 Tome $l_{11} = a_{11}$;
$u_{12} = a_{12}/l_{11}$;
$z_1 = a_{1,n+1}/l_{11}$.

Paso 2 Para $i = 2, \ldots, n-1$ tome $l_{i,i-1} = a_{i,i-1}$; (*i-ésimo renglón de L*).
$l_{ii} = a_{ii} - l_{i,i-1}u_{i-1,i}$;
$u_{i,i+1} = a_{i,i+1}/l_{ii}$; (*(i + 1)-ésima columna de U*).
$z_i = (a_{i,n+1} - l_{i,i-1}z_{i-1})/l_{ii}$

Paso 3 Tome $l_{n,n-1} = a_{n,n-1}$; (*n-ésimo renglón de L*).
$l_{nn} = a_{nn} - l_{n,n-1}u_{n-1,n}$.
$z_n = (a_{n,n+1} - l_{n,n-1}z_{n-1})/l_{nn}$.

(*Pasos 4 y 5 resuelven* $U\mathbf{x} = \mathbf{z}$.)

Paso 4 Tome $x_n = z_n$.

Paso 5 Para $i = n-1, \ldots, 1$ tome $x_i = z_i - u_{i,i+1}x_{i+1}$.

Paso 6 SALIDA $(x_1, \ldots, x_n)$;
PARAR. ■

EJEMPLO 5 Para explicar con un ejemplo el procedimiento de las matrices tridiagonales, consideremos el sistema tridiagonal de ecuaciones

$$\begin{aligned} 2x_1 - x_2 \qquad\qquad &= 1, \\ -x_1 + 2x_2 - x_3 \qquad &= 0, \\ - x_2 + 2x_3 - x_4 &= 0, \\ - x_3 + 2x_4 &= 1. \end{aligned}$$

cuya matriz aumentada es

$$\left[\begin{array}{rrrr:r} 2 & -1 & 0 & 0 & 1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & 1 \end{array}\right]$$

El algoritmo de factorización de Crout genera la factorización

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -1 & \frac{3}{2} & 0 & 0 \\ 0 & -1 & \frac{4}{3} & 0 \\ 0 & 0 & -1 & \frac{5}{4} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{2}{3} & 0 \\ 0 & 0 & 1 & -\frac{3}{4} \\ 0 & 0 & 0 & 1 \end{bmatrix} = LU.$$

Al resolver el sistema $L\mathbf{z} = \mathbf{b}$ obtenemos $\mathbf{z} = (\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 1)^t$ y la solución de $U\mathbf{x} = \mathbf{z}$ es $\mathbf{x} = (1, 1, 1, 1)^t$. ■

El algoritmo de factorización de Crout puede aplicarse siempre que $l_{ii} \neq 0$ para toda $i = 1, 2, \ldots, n$. Dos condiciones, que garantizan la veracidad de esto, son que la matriz de coeficientes del sistema sea definida positiva o que sea estrictamente dominante en sentido diagonal. Una condición adicional con que es garantiza la aplicabilidad de este algoritmo se da en el siguiente teorema, cuya demostración se incluye en el ejercicio 22.

Teorema 6.29 Supongamos que $A = (a_{ij})$ es tridiagonal con $a_{i,i-1}\, a_{i,i+1} \neq 0$, para toda $i = 2, 3, \ldots, n-1$. Si $|a_{11}| > |a_{12}|$, $|a_{ii}| \geq |a_{i,i-1}| + |a_{i,i+1}|$ para cada $i = 2, 3, \ldots, n-1$ y $|a_{nn}| > |a_{n,n-1}|$, entonces A es no singular y los valores de l_{ii} descritos en el algoritmo de factorización de Crout son distintos de cero para cada $i = 1, 2, \ldots, n$. ■

CONJUNTO DE EJERCICIOS 6.6

1. Determine cuáles de las siguientes matrices son (i) simétricas, (ii) singulares, (iii) estrictamente diagonal dominante, (iv) definidas positivas.

a. $\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$

b. $\begin{bmatrix} -2 & 1 \\ 1 & -3 \end{bmatrix}$

c. $\begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 0 \\ 1 & 0 & 4 \end{bmatrix}$

d. $\begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 1 & 2 & 4 \end{bmatrix}$

e. $\begin{bmatrix} 4 & 2 & 6 \\ 3 & 0 & 7 \\ -2 & -1 & -3 \end{bmatrix}$ **f.** $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 2 \end{bmatrix}$

g. $\begin{bmatrix} 4 & 0 & 0 & 0 \\ 6 & 7 & 0 & 0 \\ 9 & 11 & 1 & 0 \\ 5 & 4 & 1 & 1 \end{bmatrix}$ **h.** $\begin{bmatrix} 2 & 3 & 1 & 2 \\ -2 & 4 & -1 & 5 \\ 3 & 7 & 1.5 & 1 \\ 6 & -9 & 3 & 7 \end{bmatrix}$

2. Use el algoritmo de factorización LDL^t y obtenga una factorización de la forma $A = LDL^t$ para las siguientes matrices:

a. $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ **b.** $A = \begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}$

c. $A = \begin{bmatrix} 4 & 1 & -1 & 0 \\ 1 & 3 & -1 & 0 \\ -1 & -1 & 5 & 2 \\ 0 & 0 & 2 & 4 \end{bmatrix}$ **d.** $A = \begin{bmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{bmatrix}$

3. Use el algoritmo de Choleski y obtenga una factorización de la forma $A = LL^t$ para las matrices del ejercicio 2.

4. Modifique el algoritmo de factorización LDL^t como se indica en el texto, de modo que le sirva para resolver los sistemas lineales. Utilice el algoritmo modificado para resolver los siguientes sistemas lineales.

a.
$$\begin{aligned} 2x_1 - x_2 \qquad &= 3, \\ -x_1 + 2x_2 - x_3 &= -3, \\ - x_2 + 2x_3 &= 1. \end{aligned}$$

b.
$$\begin{aligned} 4x_1 + x_2 + x_3 + x_4 &= 0.65, \\ x_1 + 3x_2 - x_3 + x_4 &= 0.05, \\ x_1 - x_2 + 2x_3 \qquad &= 0, \\ x_1 + x_2 \qquad + 2x_4 &= 0.5. \end{aligned}$$

c.
$$\begin{aligned} 4x_1 + x_2 - x_3 \qquad &= 7, \\ x_1 + 3x_2 - x_3 \qquad &= 8, \\ -x_1 - x_2 + 5x_3 + 2x_4 &= -4, \\ 2x_3 + 4x_4 &= 6. \end{aligned}$$

d.
$$\begin{aligned} 6x_1 + 2x_2 + x_3 - x_4 &= 0, \\ 2x_1 + 4x_2 + x_3 \qquad &= 7, \\ x_1 + x_2 + 4x_3 - x_4 &= -1, \\ -x_1 \qquad - x_3 + 3x_4 &= -2. \end{aligned}$$

5. Modifique el algoritmo de Choleski como se sugiere en el texto, de manera que le sirva para resolver sistemas lineales, y con el algoritmo modificado resuelva los sistemas lineales del ejercicio 4.

6. Use la factorización de Crout con sistemas tridiagonales para resolver los siguientes sistemas lineales.

a.
$$\begin{aligned} x_1 - x_2 \qquad &= 0, \\ -2x_1 + 4x_2 - 2x_3 &= -1, \\ - x_2 + 2x_3 &= 1.5. \end{aligned}$$

b.
$$\begin{aligned} 3x_1 + x_2 \qquad &= -1, \\ 2x_1 + 4x_2 + x_3 &= 7, \\ 2x_2 + 5x_3 &= 9. \end{aligned}$$

c.
$$\begin{aligned} 2x_1 - x_2 \qquad &= 3, \\ -x_1 + 2x_2 - x_3 &= -3, \\ - x_2 + 2x_3 &= 1. \end{aligned}$$

d.
$$\begin{aligned}
0.5x_1 + 0.25x_2 \qquad\qquad &= 0.35,\\
0.35x_1 + 0.8x_2 + 0.4x_3 \qquad &= 0.77,\\
0.25x_2 + x_3 + 0.5x_4 &= -0.5,\\
x_3 - 2x_4 &= -2.25.
\end{aligned}$$

7. Sea A la matriz tridiagonal de 10×10 dada por $a_{ii} = 2$, $a_{i,i+1} = a_{i,i-1} = -1$, para cada $i = 2, \ldots, 9$, y en $a_{11} = a_{10,10} = 2$, $a_{12} = a_{10,9} = -1$. Sea $\mathbf{b}$ el vector columna de dimensión diez dado por $b_1 = b_{10} = 1$ y $b_i = 0$ para cada $i = 2, 3, \ldots, 9$. Resuelva $A\mathbf{x} = \mathbf{b}$ por medio de la factorización de Crout para sistemas tridiagonales.

8. Modifique la factorización LDL^t para factorizar una matriz simétrica A. [*Nota:* no siempre es posible la factorización.] Aplique el nuevo algoritmo a las siguientes matrices:

a. $A = \begin{bmatrix} 3 & -3 & 6 \\ -3 & 2 & -7 \\ 6 & -7 & 13 \end{bmatrix}$ **b.** $A = \begin{bmatrix} 3 & -6 & 9 \\ -6 & 14 & -20 \\ 9 & -20 & 29 \end{bmatrix}$

c. $A = \begin{bmatrix} -1 & 2 & 0 & 1 \\ 2 & -3 & 2 & -1 \\ 0 & 2 & 5 & 6 \\ 1 & -1 & 6 & 12 \end{bmatrix}$ **d.** $A = \begin{bmatrix} 2 & -2 & 4 & -4 \\ -2 & 3 & -4 & 5 \\ 4 & -4 & 10 & -10 \\ -4 & 5 & -10 & 14 \end{bmatrix}$

9. ¿Cuáles de las matrices simétricas del ejercicio 8 son definidas positivas?

10. Obtenga α de modo que $A = \begin{bmatrix} \alpha & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 4 \end{bmatrix}$ sea definida positiva.

11. Obtenga α de modo que $A = \begin{bmatrix} 2 & \alpha & -1 \\ \alpha & 2 & 1 \\ -1 & 1 & 4 \end{bmatrix}$ sea definida positiva.

12. Obtenga α y $\beta > 0$ de modo que la matriz

$$A = \begin{bmatrix} 4 & \alpha & 1 \\ 2\beta & 5 & 4 \\ \beta & 2 & \alpha \end{bmatrix}$$

sea estrictamente diagonal dominante.

13. Obtenga $\alpha > 0$ y $\beta > 0$ de modo que la matriz

$$A = \begin{bmatrix} 3 & 2 & \beta \\ \alpha & 5 & \beta \\ 2 & 1 & \alpha \end{bmatrix}$$

es estrictamente dominante en sentido diagonal.

14. Suponga que A y B son matrices de $n \times n$ estrictamente diagonal dominante.

a. ¿Es $-A$ estrictamente diagonal dominante?

b. ¿Es A^t estrictamente diagonal dominante?

c. ¿Es $A + B$ estrictamente diagonal dominante?

d. ¿Es A^2 estrictamente diagonal dominante?

e. ¿Es $A - B$ estrictamente diagonal dominante?

15. Suponga que A y B son matrices definidas positivas de $n \times n$.

a. ¿Es $-A$ definida positiva?

b. ¿Es A^t definida positiva?

c. ¿Es $A + B$ definida positiva?

d. ¿Es A^2 definida positiva?

e. ¿Es $A - B$ definida positiva?

16. Sea

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ -1 & 1 & \alpha \end{bmatrix}.$$

Calcule todos los valores de α para los cuales

a. A es singular.

b. A es estrictamente diagonal dominante.

c. A es simétrica.

d. A es definida positiva.

17. Sea

$$A = \begin{bmatrix} \alpha & 1 & 0 \\ \beta & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

Obtenga los valores de α y de β para los cuales

a. A es singular.

b. A es estrictamente diagonal dominante.

c. A es simétrica.

d. A es definida positiva.

18. Suponga que A y B conmutan, es decir, $AB = BA$. ¿Deben conmutar también A^t y B^t?

19. Construya una matriz A que no sea simétrica, pero para la cual $\mathbf{x}^t A \mathbf{x} > 0$ para toda $\mathbf{x} \neq 0$.

20. Demuestre que puede efectuarse la eliminación gaussiana en A sin intercambios de renglones, si y sólo si todas las primeras submatrices principales de A son no singulares. [*Sugerencia:* divida todas las matrices de la ecuación

$$A^{(k)} = M^{(k-1)}M^{(k-2)} \ldots M^{(1)} A$$

entre las columnas k-ésima y $(k + 1)$-ésima y horizontalmente entre los renglones k-ésimo y $(k + 1)$-ésimo (véase el ejercicio 10 de la sección 6.3). Demuestre que la no singularidad de la primera submatriz principal de A equivale a $a_{kk}^{(k)} \neq 0$.]

21. Las matrices tridiagonales suelen marcarse mediante la notación

$$A = \begin{bmatrix} a_1 & c_1 & 0 & \cdots & \cdots & 0 \\ b_2 & a_2 & c_2 & \ddots & & \vdots \\ 0 & b_3 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & \cdots & 0 & b_n & a_n \end{bmatrix}$$

para subrayar que no es necesario tener en cuenta todos sus elementos. Reescriba el algoritmo de factorización de Crout utilizando esta notación y modifique la notación de l_{ij} y de u_{ij} en una forma similar.

22. Demuestre el teorema 6.29. [*Sugerencia*: demuestre que $|u_{i,i+1}| < 1$ para cada $i = 1, 2, \ldots, n-1$ y que $|l_{ii}| > 0$ para cada $i = 1, 2, \ldots, n$. Deduzca que $\det A = \det L \cdot \det U \neq 0$.]

23. Suponga que $V = 5.5$ volts en el primer ejemplo de este capítulo. Al reordenar las ecuaciones, podemos formar un sistema lineal tridiagonal. Use el algoritmo de factorización de Crout para obtener la solución del sistema modificado.

24. Use el algoritmo de factorización de Crout para construir el conteo de operaciones para resolver un sistema lineal de $n \times n$.

25. En un trabajo de Dorn y Burdick [DB] se indica lo siguiente: la longitud promedio de las alas que resulta de aparear tres variedades mutantes de moscas de fruta (*Drosophila melanogaster*) puede expresarse en forma de la matriz simétrica

$$A = \begin{bmatrix} 1.59 & 1.69 & 2.13 \\ 1.69 & 1.31 & 1.72 \\ 2.13 & 1.72 & 1.85 \end{bmatrix},$$

donde a_{ij}, denota la longitud promedio de las alas de una cría obtenida al aparear un macho de tipo i con una hembra de tipo j.

a. ¿Cuál es la importancia física de la simetría de esta matriz?

b. ¿Es la matriz definida positiva? De ser así, demuéstrelo; en caso contrario, obtenga un vector $\mathbf{x}$ distinto de cero para el cual $\mathbf{x}^t A\mathbf{x} \leq 0$.

26. Suponga que la matriz definida positiva A tiene la factorización de Choleski $A = LL^t$ y también la factorización $A = \hat{L}D\hat{L}^t$, donde D es la matriz diagonal con elementos positivos diagonales $d_{11}, d_{22}, \ldots, d_{nn}$. Sea $D^{1/2}$ la matriz diagonal con elementos diagonales $\sqrt{d_{11}}, \sqrt{d_{22}}, \ldots, \sqrt{d_{nn}}$.

a. Demuestre que $D = D^{1/2}D^{1/2}$.

b. Demuestre que $L = \hat{L}D^{1/2}$.

6.7 Reseña de métodos y de software

En este capítulo estudiamos los métodos directos para resolver los sistemas lineales. Un sistema lineal consta de n ecuaciones con n incógnitas expresadas en la notación matricial como $A\mathbf{x} = \mathbf{b}$. Estas técnicas utilizan una serie finita de operaciones aritméticas para determinar la solución exacta del sistema, sujeta únicamente al error de redondeo. Descubrimos que el sistema lineal $A\mathbf{x} = \mathbf{b}$ tiene una solución única si y sólo si existe A^{-1} lo cual equivale a $\det A \neq 0$. La solución del sistema lineal es el vector $\mathbf{x} = A^{-1}\mathbf{b}$.

Explicamos las técnicas de pivoteo para reducir al mínimo los efectos del error de redondeo, que pueden dominar la solución cuando se emplean métodos directos. Describimos el pivoteo parcial, el pivoteo parcial escalado y el pivoteo total. Recomendamos los dos primeros procedimientos para resolver la generalidad de los problemas, pues ambos disminuyen los efectos del error de redondeo sin aumentar excesivamente los cálculos adicionales. El pivoteo total deberá emplearse si se sospecha que el error de redondeo es grande. En la sección 7.4 veremos algunos procedimientos para estimar este error.

Demostramos que, con ligeras modificaciones, la eliminación gaussiana produce una factorización de la matriz A en LU, donde L es una matriz triangular inferior con unos en la diagonal y donde U es una matriz triangular superior. A este proceso se le llama factorización de Doolittle. No todas las matrices no singulares pueden ser factorizadas en esta forma, pero una permutación de los renglones dará siempre una factorización de la forma $PA = LU$, donde P es la matriz de permutación con que se rearreglan los renglones de A. La ventaja de la factorización consiste en que el trabajo se reduce cuando se resuelven sistemas lineales $A\mathbf{x} = \mathbf{b}$ con la misma matriz de coeficientes A y con diferentes vectores $\mathbf{b}$.

Cuando la matriz A es definida positiva, las factorizaciones adoptan una forma más simple. Por ejemplo, la factorización de Choleski presenta la forma $A = LL^t$, donde L es una matriz triangular inferior. Las matrices definidas positivas también pueden factorizarse en la forma $A = LDL^t$ donde L es una matriz triangular inferior con unos en la diagonal y donde D es la diagonal. Con estas factorizaciones, es posible simplificar el manejo de A. Si A es tridiagonal, la factorización LU ofrecerá una forma muy simple: L tiene unos en la diagonal principal y 0 en las demás, salvo en la diagonal situada inmediatamente debajo de la diagonal principal. Además, U tiene exclusivamente elementos distintos de cero en la diagonal principal y en la diagonal situada por arriba

Los métodos directos son los procedimientos de elección en la generalidad de los sistemas lineales. Los métodos especiales se recomiendan en el caso de matrices tridiagonales, con banda y definidas positivas. En el caso general, se recomienda la eliminación gaussiana o los métodos de factorización LU, que permiten el pivoteo. En estos casos, conviene vigilar los efectos del error de redondeo. En la sección 7.4 veremos cómo estimar los errores de los métodos directos.

Los sistemas lineales grandes, en los cuales ocurren fundamentalmente elementos cero en patrones regulares, pueden resolverse de manera eficiente aplicando un procedimiento iterativo como los que se exponen en el capítulo 7. Este tipo de sistemas se presenta, por ejemplo, cuando se emplean técnicas de diferencia finita para resolver problemas con valor de frontera, aplicación muy común en la solución numérica de las ecuaciones diferenciales parciales.

A veces es muy difícil resolver un sistema lineal grande que tiene básicamente elementos distintos de cero o cuando los elementos cero no muestran un patrón predecible. La matriz asociada al sistema puede colocarse en un almacenamiento secundario en la forma particionada y las partes se introducen en la memoria principal sólo cuando se requieren para los cálculos. Los métodos que requieren almacenamiento secundario pueden ser iterativos o directos, pero generalmente exigen técnicas de las estructuras de datos y la teoría de gráficas. Al lector que desee profundizar en el tema le recomendamos consultar [BuR] y [RW], donde encontrará una amplia explicación de las técnicas actuales.

Los programas para realizar operaciones con matrices y la solución directa de los sistemas lineales implantados en IMSL y NAG se basan en LAPACK, paquete de subrutinas de dominio público. Este paquete contiene una excelente documentación que también puede obtenerse de libros acerca de él. Concentraremos nuestra atención en varias de las subrutinas que vienen en las tres fuentes anteriores.

LAPACK trae un conjunto de operaciones de nivel inferior, denominado Basic Linear Algebra Subprograms (BLAS). EL nivel 1 de BLAS generalmente contiene operaciones vectoriales con datos de entrada y conteos de operaciones de $O(n)$. El nivel 2 consta de operaciones de matrices y vectores con datos de entrada y conteo de operaciones $O(n^2)$. El nivel 3 consta de las operaciones matriciales con datos de entrada y conteos de operaciones $O(n^3)$. Por ejemplo, en el nivel 1 la subrutina SCOPY sobrescribe un vector $\mathbf{y}$ con un vector $\mathbf{x}$; SSCAL calcula a veces un escalar por un vector $\mathbf{x}$; SAXPY agrega a un vector un escalar multiplicado por un vector ($\mathbf{y} = \mathbf{a} \cdot \mathbf{x} + \mathbf{y}$); SDOT calcula el producto interno o escalar de dos vectores; SNRM2 calcula la norma euclidiana de un vector aplicando un método semejante al de la sección 1.4, ISAMAX calcula el índice de un componente vectorial que da el valor máximo absoluto de todos los componentes. En el nivel 2, SGEMV calcula el producto de una matriz y un vector y el nivel 3, SGEMM calcula el producto de una matriz y un vector.

Las subrutinas de LAPACK con que se resuelven los sistemas lineales factorizan primero la matriz A. La factorización depende del tipo de matriz, en la siguiente forma:

1. Matriz general $PA = LU$;
2. Matriz definida positiva $A = LL^t$;
3. Matriz simétrica $A = LDL^t$;
4. Matriz tridiagonal $A = LU$ (en forma de banda).

La subrutina STRTRS resuelve un sistema lineal triangular cuando la matriz es triangular inferior o superior.

La subrutina SGETRF factoriza PA en LU como operación preliminar de la subrutina SGETRS, la cual calcula después la solución de $A\mathbf{x} = \mathbf{b}$. La subrutina SGETRI sirve para construir la inversa de una matriz A y para calcular la determinante de A una vez que ésta fue factorizada por medio de SGETRF.

La factorización de Choleski para una matriz definida positiva A se obtiene mediante la subrutina SPOTRF. El sistema lineal $A\mathbf{x} = \mathbf{b}$ puede resolverse, entonces, aplicando la subrutina SPOTRS. Con SPOTRI pueden calcularse las inversas y los determinantes de las matrices definidas positivas, cuando se utiliza la factorización de Choleski. Si A es simétrica, la factorización LDL^t se obtiene aplicando SSYTRF. Después, los sistemas lineales pueden resolverse por medio de SSYTRS. Si se desean inversas o determinantes, puede utilizarse SSYTRI.

Muchas de las subrutinas de LINPACK, y su sucesor LAPACK, pueden ejecutarse por medio de MATLAB. Una matriz no singular A se factoriza aplicando el comando

$$[L, U, P] = lu(A)$$

en la forma $PA = LU$, donde P es la matriz de permutación que se define al realizar el pivoteo parcial para resolver un sistema lineal que contenga A. Si en MATLAB se definieron la matriz no singular A y el vector $\mathbf{b}$, el comando

$$x = A \setminus b$$

resuelve el sistema lineal usando primero el comando de factorización $PA = LU$. Después resuelve el sistema triangular inferior $L\mathbf{z} = \mathbf{b}$ para $\mathbf{z}$ por medio del comando,

$$z = L \setminus b$$

Esto se acompaña de una solución del sistema triangular superior $U\mathbf{x} = \mathbf{z}$ aplicando el comando

$$x = U \setminus z$$

Otros comandos de MATLAB incluyen el cálculo de la inversa, la transpuesta y el determinante de la matriz A por medio de los comandos $inv(A)$, A' y $\det(A)$, respectivamente.

La biblioteca IMSL incluye equivalentes de casi todas las subrutinas de LAPACK y además algunas extensiones. Se les asigna un nombre que indica las tareas que realizan, a saber:

1. Se usan las tres primeras letras del nombre.
   - **a.** LSL: resuelve un sistema lineal
   - **b.** LFT: factoriza una matriz de coeficientes
   - **c.** LFS: resuelve un sistema lineal dados los factores de LFT
   - **d.** LFD: calcula los determinantes de los factores dados
   - **e.** LIN: calcula la inversa de los factores dados

2. Las dos últimas letras indican el tipo de matriz en cuestión.
   a. RG: real general
   b. RT: real triangular
   c. DS: real definida positiva
   d. SF: real simétrica
   e. RB: real con banda

Por ejemplo, la rutina LFTDS factoriza una matriz real positiva definida.

La biblioteca NAG cuenta con muchas subrutinas de métodos directos para resolver sistemas lineales, las cuales se asemejan a las de LAPACK y de IMSL. Por ejemplo, la subrutina F04AEF resuelve sistemas lineales mediante la factorización de Crout. La subrutina F04ATF resuelve un solo sistema lineal mediante la factorización de Crout, igual que F04AEF. La subrutina F04EAF resuelve un solo sistema lineal en el que la matriz es real y tridiagonal, y F04ASF lo resuelve cuando la matriz es real y positiva definida. Las matrices inversas pueden calcularse por medio de F07AJF después de usar F07ADF para una matriz real y para F01ABF en el caso de una matriz definida positiva. Un determinante puede calcularse con F03AAF. La factorización puede obtenerse usando F07ADF en la factorización *LU* de una matriz real y usando F01LEF con una matriz tridiagonal; los sistemas lineales pueden resolver, entonces, mediante F07AEF. La factorización de Choleski para una matriz definida positiva puede obtenerse aplicando F07FDF, y un sistema lineal puede obtenerse aplicando F07FEF. La biblioteca NAG también incluye manejo de matriz y vectores de nivel inferior.

En Golub y Van Loan [GV], en Forsythe y Moler [FM] y en Stewart [St], se puede encontrar información complementaria sobre la solución numérica de sistemas lineales y de matrices. El empleo de métodos directos para resolver grandes sistemas analíticos se explica ampliamente en George y Liu [GL] y en Pissanetzky [Pi]. Coleman y Van Loan [CV] estudian la utilización de BLAS, LINPACK y MATLAB.

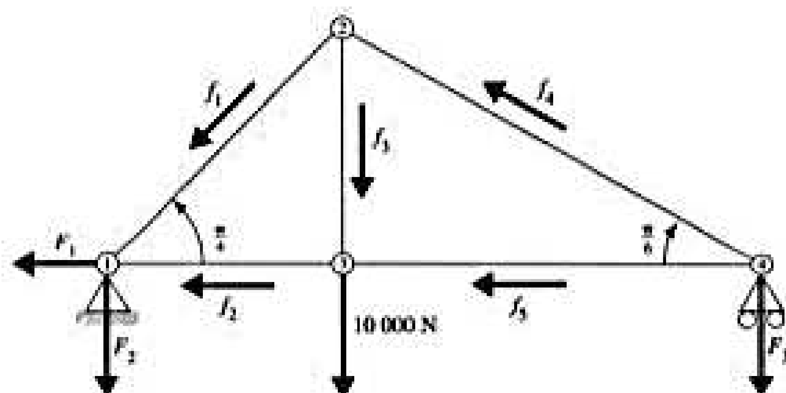
CAPÍTULO

7

Métodos iterativos en el álgebra matricial

. . .

Los armazones son estructuras ligeras capaces de soportar cargas pesadas. En el diseño de puentes, los miembros de la estructura están conectados con juntas rotatorias de pasador que permiten transferir las fuerzas de un miembro a otro. La figura anexa muestra una estructura que se mantiene estacionaria en el extremo inferior de la izquierda ①, que se desplaza horizontalmente en el extremo inferior derecho ④ y que tiene juntas de pasador en ①, ②, ③ y ④. Se coloca una carga de 10 000 newtons (N) en la junta ③ y las fuerzas que actúan sobre los miembros de la estructura tienen magnitudes dadas por f_1, f_2, f_3, f_4 y f_5 como se observa en la figura. El miembro de soporte estacionario tiene una fuerza horizontal F_1 y una fuerza vertical F_2, pero el miembro de soporte movible tiene únicamente la fuerza vertical F_3.



Si la estructura está en equilibrio estático, las fuerzas en cada junta deben agregarse al vector cero, de modo que la suma de las componentes horizontal y vertical de las fuerzas en cada junta debe ser cero. Esto genera el sistema de ecuaciones lineales que aparecen

en la tabla adjunta. Una matriz de 8 × 8 que lo describe tiene 46 elementos cero y apenas 18 elementos no cero. Las matrices con un alto porcentaje de elementos cero reciben el nombre de *esparcidas* y, a menudo, se resuelven aplicando métodos iterativos más que directos. La solución iterativa de este sistema se incluye en el ejercicio 18 de la sección 7.3.

Junta	Componentes horizontales	Componentes verticales
①	$-F_1 + \frac{\sqrt{2}}{2} f_1 + f_2 = 0$	$\frac{\sqrt{2}}{2} f_1 - F_2 = 0$
②	$-\frac{\sqrt{2}}{2} f_1 + \frac{\sqrt{3}}{2} f_4 = 0$	$-\frac{\sqrt{2}}{2} f_1 - f_3 + \frac{1}{2} f_4 = 0$
③	$-f_2 + f_5 = 0$	$f_3 - 10\,000 = 0$
④	$-\frac{\sqrt{3}}{2} f_4 - f_5 = 0$	$\frac{1}{2} f_4 - F_3 = 0$

Los métodos estudiados en el capítulo 6 usaron métodos directos para resolver un sistema de $n \times n$ ecuaciones lineales de la forma $A\mathbf{x} = \mathbf{b}$. En este capítulo estudiaremos los métodos iterativos con que se resuelve un sistema de este tipo.

7.1 Normas de vectores y de matrices

En el capítulo 2 explicamos los métodos iterativos para obtener las raíces de ecuaciones de la forma $f(x) = 0$. Encontramos una aproximación inicial (o aproximaciones), y luego determinamos otras nuevas basándonos en qué tan bien las aproximaciones anteriores satisfacían la ecuación. Antes de empezar el estudio de los métodos iterativos con que se resuelven los sistemas lineales, necesitamos contar primero con un medio que nos permita medir la distancia entre los vectores columna n-dimensionales, para determinar si una serie de esos vectores convergen a una solución del sistema. En la práctica, también necesitamos esta medida cuando la solución se obtiene por los métodos directos expuestos en el capítulo 6. Estos requirieron muchas operaciones aritméticas, y el uso de la aritmética de dígitos finitos conduce sólo a una aproximación de la solución real del sistema.

Denotemos con $\mathbb{R}^n$ el conjunto de todos los vectores columna n-dimensionales con componentes de números reales. Si queremos definir una distancia en $\mathbb{R}^n$ utilizaremos la noción de una norma.

Definición 7.1 Una **norma vectorial** en $\mathbb{R}^n$ es una función, $\|\cdot\|$, de $\mathbb{R}^n$ a $\mathbb{R}$ con las siguientes propiedades:

(i) $\|\mathbf{x}\| \geq 0$ para toda $\mathbf{x} \in \mathbb{R}^n$,

(ii) $\|\mathbf{x}\| = 0$ si y sólo si $\mathbf{x} = \mathbf{0}$,

(iii) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ para todo $\alpha \in \mathbb{R}$ y $\mathbf{x} \in \mathbb{R}^n$,

(iv) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ para todo $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. ■

Necesitaremos sólo dos normas específicas en $\mathbb{R}^n$ aunque en el ejercicio 2 se presenta una tercera norma de este vector.

Como los vectores de $\mathbb{R}^n$ son vectores columna, conviene utilizar la notación de la transpuesta que se vio en la sección 6.3, cuando representamos un vector en función de sus componentes. Por ejemplo, el vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

se escribirá $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$.

Definición 7.2 Las normas l_2 y l_∞ del vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ están definidas por

$$\|\mathbf{x}\|_2 = \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2} \quad \text{y} \quad \|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i|.$$ ■

La norma l_2 se llama **norma euclideana** del vector $\mathbf{x}$, dado que representa la noción común de distancia respecto al origen en caso de que $\mathbf{x}$ esté en $\mathbb{R}^1 \equiv \mathbb{R}$, $\mathbb{R}^2$, o bien $\mathbb{R}^3$. Por ejemplo, la norma l_2, del vector $\mathbf{x} = (x_1, x_2, x_3)^t$ denota la longitud del segmento de recta que une los puntos $(0, 0, 0)$ y (x_1, x_2, x_3). En la figura 7.1 se muestra la frontera de los vectores en $\mathbb{R}^2$ y $\mathbb{R}^3$ que tienen la norma l_2 menor que 1. La figura 7.2 contiene un ejemplo semejante de la norma l_∞.

Figura 7.1

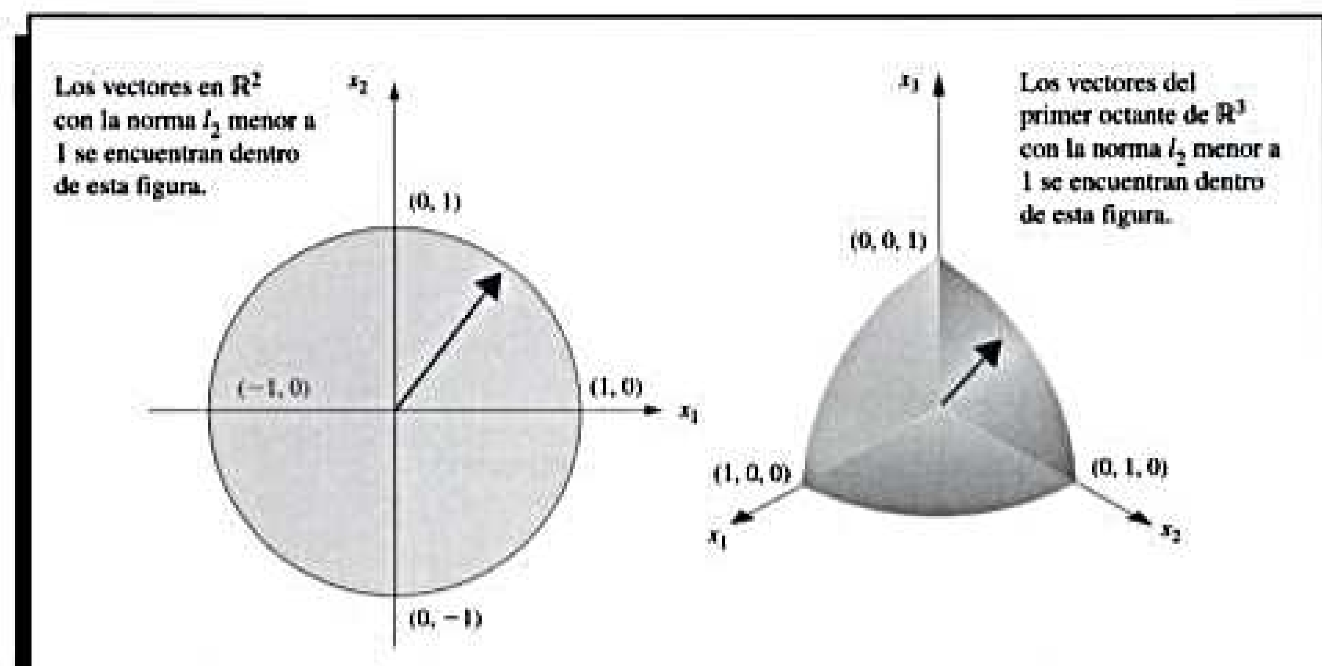
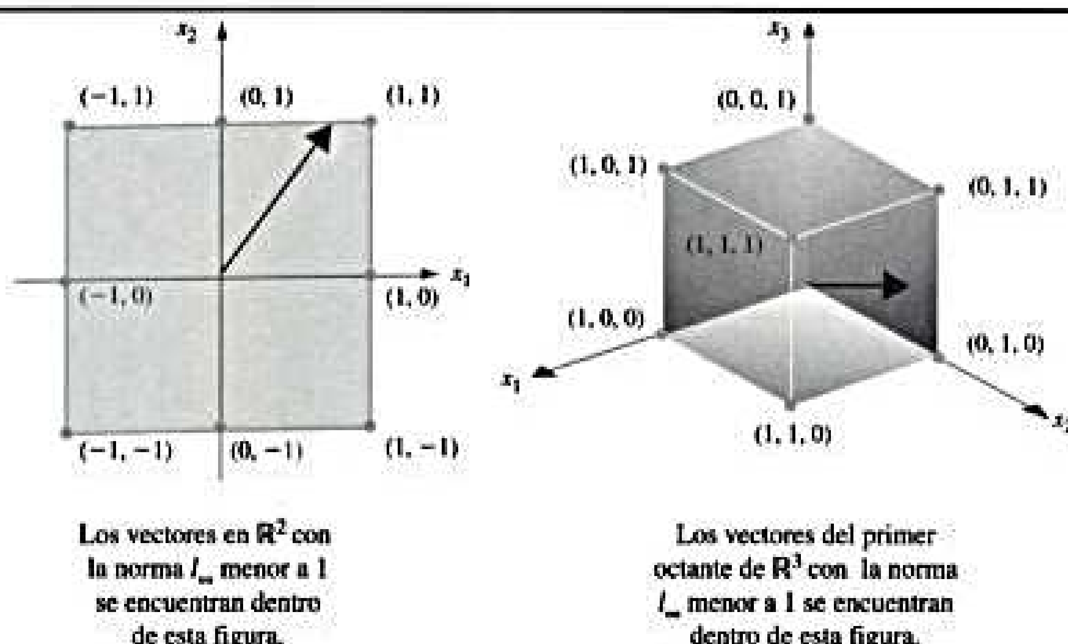


Figura 7.2



EJEMPLO 1 El vector $\mathbf{x} = (-1, 1, -2)^t$ en $\mathbb{R}^3$ tiene las normas

$$\|\mathbf{x}\|_2 = \sqrt{(-1)^2 + (1)^2 + (-2)^2} = \sqrt{6}$$

y

$$\|\mathbf{x}\|_\infty = \text{máx}\,(|-1|, |1|, |-2|) = 2.$$

■

Es fácil demostrar que las propiedades de la definición 7.1 se aplican también a la norma l_∞ pues provienen de resultados semejantes para valores absolutos. Por ejemplo, si $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$, entonces

$$\|\mathbf{x} + \mathbf{y}\|_\infty = \max_{1\le i\le n} |x_i + y_i| \le \max_{1\le i\le n} (|x_i| + |y_i|) \le \max_{1\le i\le n} |x_i| + \max_{1\le i\le n} |y_i| = \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty.$$

Si queremos demostrar que

$$\|\mathbf{x} + \mathbf{y}\|_2 \le \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2, \quad \text{para todo } \mathbf{x}, \mathbf{y} \in \mathbb{R}_n,$$

necesitamos una famosa desigualdad.

Teorema 7.3 **(Desigualdad de Cauchy-Buniakowsky-Schwarz)**

Para toda $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$, en $\mathbb{R}^n$,

$$\mathbf{x}^t\,\mathbf{y} = \sum_{i=1}^{n} x_i\,y_i \le \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2} \left\{\sum_{i=1}^{n} y_i^2\right\}^{1/2} = \|\mathbf{x}\| \cdot \|\mathbf{y}\|. \qquad (7.1)$$

■

Demostración Si $\mathbf{y} = \mathbf{0}$ o $\mathbf{x} = \mathbf{0}$, el resultado será inmediato porque ambos lados de la desigualdad son cero.

Supongamos que $\mathbf{y} \neq \mathbf{0}$ y que $\mathbf{x} \neq \mathbf{0}$. Para cada $\lambda \in \mathbb{R}$,

$$0 \leq \|\mathbf{x} - \lambda\mathbf{y}\|_2^2 = \sum_{i=1}^{n} (x_i - \lambda y_i)^2 = \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i y_i + \lambda^2 \sum_{i=1}^{n} y_i^2,$$

así que

$$2\lambda \sum_{i=1}^{n} x_i y_i \leq \sum_{i=1}^{n} x_i^2 + \lambda^2 \sum_{i=1}^{n} y_i^2 = \|\mathbf{x}\|_2^2 + \lambda^2\|\mathbf{y}\|_2^2.$$

Como $\|\mathbf{x}\|_2 > 0$ y $\|\mathbf{y}\|_2 > 0$, podemos hacer $\lambda = \|\mathbf{x}\|_2/\|\mathbf{y}\|_2$ para obtener

$$\left(2\frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2}\right)\left(\sum_{i=1}^{n} x_i y_i\right) \leq \|\mathbf{x}\|_2^2 + \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{y}\|_2^2}\|\mathbf{y}\|_2^2 = 2\|\mathbf{x}\|_2^2,$$

así

$$2\sum_{i=1}^{n} x_i y_i \leq 2\|\mathbf{x}\|_2^2 \frac{\|\mathbf{y}\|_2}{\|\mathbf{x}\|_2} = 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2.$$

Por tanto,

$$\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{n} x_i y_i \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2 = \left\{\sum_{i=1}^{n} x_i^2\right\}^{1/2}\left\{\sum_{i=1}^{n} y_i^2\right\}^{1/2}$$ ■ ■ ■

Con este resultado vemos que para cada $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{x} + \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \leq \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2,$$

lo cual nos da la última propiedad de la norma

$$\|\mathbf{x} + \mathbf{y}\|_2 \leq (\|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{x}\|_2^2)^{1/2} = \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2.$$

La norma de un vector proporciona una medida de la distancia entre un vector arbitrario y el vector cero; por ello, podemos definir la **distancia entre dos vectores** como la norma de la diferencia de los vectores.

Definición 7.4 Si $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ son vectores en $\mathbb{R}^n$, las distancias l_2 y l_∞ entre $\mathbf{x}$ y $\mathbf{y}$ están definidas por

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left\{\sum_{i=1}^{n} (x_i - y_i)^2\right\}^{1/2} \quad \text{y} \quad \|\mathbf{x} - \mathbf{y}\|_\infty = \max_{1\leq i\leq n} |x_i - y_i|.$$ ■

EJEMPLO 2 El sistema lineal

$$\begin{aligned}3.3330x_1 + 15920x_2 - 10.333x_3 &= 15913,\\ 2.2220x_1 + 16.710x_2 + 9.6120x_3 &= 28.544,\\ 1.5611x_1 + 5.1791x_2 + 1.6852x_3 &= 8.4254\end{aligned}$$

tiene la solución $(x_1, x_2, x_3)^t = (1, 1, 1)^t$. Si efectuamos la eliminación gaussiana en la aritmética de redondeo a cinco dígitos utilizando el pivoteo parcial (algoritmo 6.2), la solución que se obtiene es

$$\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^t = (1.2001, 0.99991, 0.92538)^t.$$

Las mediciones de $\mathbf{x} - \tilde{\mathbf{x}}$ están dadas por

$$\begin{aligned}\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty &= \text{máx}\{|1 - 1.2001|, |1 - 0.99991|, |1 - 0.92538|\}\\ &= \text{máx}\{0.2001, 0.00009, 0.07462\} = 0.2001\end{aligned}$$

y por

$$\begin{aligned}\|\mathbf{x} - \tilde{\mathbf{x}}\|_2 &= [(1 - 1.2001)^2 + (1 - 0.99991)^2 + (1 - 0.92538)^2]^{1/2}\\ &= [(0.2001)^2 + (0.00009)^2 + (0.07462)^2]^{1/2} = 0.21356.\end{aligned}$$

Aunque las componentes $\tilde{x}_2$ y $\tilde{x}_3$ son buenas aproximaciones de x_2 y de x_3, la componente $\tilde{x}_1$ es una aproximación deficiente de x_1, y $|x_1 - \tilde{x}_1|$ domina las normas. ■

El concepto de distancia en $\mathbb{R}^n$ también sirve para definir el límite de una sucesión de vectores en este espacio.

Definición 7.5 Se dice que una sucesión $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ de vectores en $\mathbb{R}^n$ **converge** a $\mathbf{x}$ respecto a la norma $\|\cdot\|$ si dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| < \varepsilon, \quad \text{para todo } k \geq N(\varepsilon). \qquad ■$$

Teorema 7.6 La sucesión de vectores $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$ en $\mathbb{R}^n$ respecto a $\|\cdot\|_\infty$ si y sólo si $\lim_{k\to\infty} x_i^{(k)} = x_i$, para cada $i = 1, 2, \ldots, n$. ■

Demostración Supongamos que $\{\mathbf{x}^{(k)}\}$ converge en $\mathbf{x}$ respecto a $\|\cdot\|_\infty$. Dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que para toda $k \geq N(\varepsilon)$,

$$\max_{i=1,2,\ldots,n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \varepsilon.$$

Este resultado implica que $|x_i^{(k)} - x_i| < \varepsilon$, para cada $i = 1, 2, \ldots, n$, de modo que para cada i, $\lim_{k\to\infty} x_i^{(k)} = x_i$.

Por el contrario, supongamos que $\lim_{k\to\infty} x_i^{(k)} = x_i$, para cada $i = 1, 2, \ldots, n$. Para cualquier $\varepsilon > 0$, sea $N_i(\varepsilon)$, para cada i, la representación de un entero con la propiedad de que

$$|x_i^{(k)} - x_i| < \varepsilon,$$

siempre que $k \geq N_i(\varepsilon)$.

Definimos $N(\varepsilon) = \max_{i=1,2,\ldots,n} N_i(\varepsilon)$. Si $k \geq N(\varepsilon)$, entonces

$$\max_{i=1,2,\ldots,n} |x_i^{(k)} - x_i| = \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \varepsilon.$$

Esto implica que $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$ con respecto a $\|\cdot\|_\infty$. ■ ■ ■

EJEMPLO 3 Definamos $\mathbf{x}^{(k)} \in \mathbb{R}^4$ como

$$\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t = \left(1, 2 + \frac{1}{k}, \frac{1}{k^2}, e^{-k} \operatorname{sen} k\right)^t.$$

Puesto que $\lim_{k\to\infty} 1 = 1$, $\lim_{k\to\infty}(2 + 1/k) = 2$, $\lim_{k\to\infty} 3/k^2 = 0$ y $\lim_{k\to\infty} e^{-k} \operatorname{sen} k = 0$. El teorema 7.6 implica que la sucesión $\{\mathbf{x}^{(k)}\}$ converge a $(1, 2, 0, 0)^t$ respecto a $\|\cdot\|_\infty$. ■

Es muy complicado demostrar directamente que la sucesión del ejemplo 3 converge a $(1, 2, 0, 0)^t$ respecto a la norma l_2. Resulta más fácil probar el siguiente resultado y aplicarlo después a este caso especial.

Teorema 7.7 Para todo $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\, \|\mathbf{x}\|_\infty.$$

■

Demostración Sea x_j la coordenada de $\mathbf{x}$ tal que $\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i| = |x_j|$. Entonces

$$\|\mathbf{x}\|_\infty^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 = \|\mathbf{x}\|_2^2,$$

por tanto

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2,$$

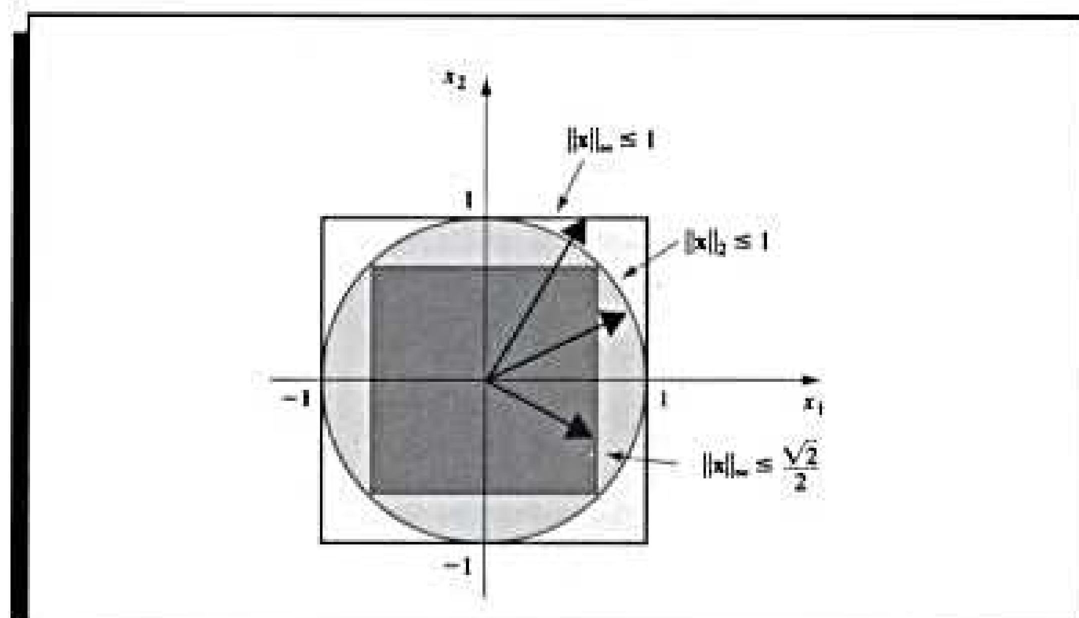
y

$$\|\mathbf{x}\|_2^2 \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = n x_j^2 = n\|\mathbf{x}\|_\infty^2,$$

por tanto $\|\mathbf{x}\|_2 \leq \sqrt{n}\, \|\mathbf{x}\|_\infty$. ■ ■ ■

La figura 7.3 muestra gráficamente este resultado cuando $n = 2$.

Figura 7.3



EJEMPLO 4 En el ejemplo 3 vimos que la sucesión $\{\mathbf{x}^{(k)}\}$, definida por

$$\mathbf{x}^{(k)} = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k}\operatorname{sen} k\right)^t,$$

converge a $\mathbf{x} = (1, 2, 0, 0)^t$ respecto a $\|\cdot\|_\infty$. Dada cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon/2)$ con la propiedad de que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < \frac{\varepsilon}{2},$$

siempre que $k \ge N(\varepsilon/2)$. Conforme al teorema 7.7, lo anterior implica que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 < \sqrt{4}\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < 2(\varepsilon/2) = \varepsilon,$$

cuando $k \ge N(\varepsilon/2)$. Por tanto, $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$ respecto a $\|\cdot\|_2$. ■

Puede demostrarse que todas las normas de $\mathbb{R}^n$ son equivalentes respecto a la convergencia; es decir, si $\|\cdot\|$ y $\|\cdot\|'$ son dos normas cualesquiera en $\mathbb{R}^n$ y si $\{\mathbf{x}^{(k)}\}_{k=1}^\infty$ tiene el límite $\mathbf{x}$ respecto a $\|\cdot\|$ entonces $\{\mathbf{x}^{(k)}\}_{k=1}^\infty$ también tendrá el límite $\mathbf{x}$ respecto a $\|\cdot\|'$. La demostración de este hecho para el caso general se encuentra en [Or2, p. 8]. Del teorema 7.7 se deduce el caso de las normas $\|\cdot\|_2$ y $\|\cdot\|_\infty$.

En las secciones siguientes de este capítulo, y de otros posteriores, necesitaremos métodos para determinar la distancia entre las matrices de $n \times n$. Para ello se requiere una vez más el concepto de norma.

Definición 7.8 Una **norma matricial** sobre el conjunto de todas las matrices de $n \times n$ es una función de valor real, $\|\cdot\|$, definida en este conjunto y que satisface para todas las matrices A y B de $n \times n$ y todos los números reales α:

(i) $\|A\| \ge 0$.
(ii) $\|A\| = 0$, si y sólo si A es O, la matriz con todas las entradas cero.

(iii) $\|\alpha A\| = |\alpha|\,\|A\|$.
(iv) $\|A + B\| \leq \|A\| + \|B\|$.
(v) $\|AB\| \leq \|A\|\,\|B\|$. ■

La **distancia entre las matrices** A y B de $n \times n$ respecto a esta norma matricial es $\|A - B\|$.

Aunque las normas matriciales se pueden obtener de varias formas, aquí únicamente consideraremos las que son consecuencia natural de las normas vectoriales l_2 y l_∞.

El siguiente teorema no es difícil de demostrar, y su demostración se deja para el ejercicio 13.

Teorema 7.9 Si $\|\cdot\|$ es una norma vectorial de $\mathbb{R}^n$, entonces

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

es una norma matricial. ■

A ésta se le llama **norma matricial natural** o *inducida* asociadada con la norma vectorial. En este libro supondremos que todas las normas matriciales son naturales, a menos que especificamos lo contrario.

Para cualquier $\mathbf{z} \neq \mathbf{0}$, tenemos que $\mathbf{x} = \mathbf{z}/\|\mathbf{z}\|$ es un vector unitario. Por lo tanto,

$$\max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \max_{\mathbf{z}\neq\mathbf{0}} \left\|A\left(\frac{\mathbf{z}}{\|\mathbf{z}\|}\right)\right\| = \max_{\mathbf{z}\neq\mathbf{0}} \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|},$$

y podemos escribir en forma alternativa

$$\|A\| = \max_{\mathbf{z}\neq\mathbf{0}} \frac{\|A\mathbf{z}\|}{\|\mathbf{z}\|}. \tag{7.2}$$

El siguiente corolario del teorema 7.9 es consecuencia de esta representación de $\|A\|$.

Corolario 7.10 Para todo vector $\mathbf{z} \neq \mathbf{0}$, matriz A y cualquier norma natural $\|\cdot\|$, tenemos

$$\|A\mathbf{z}\| \leq \|A\| \cdot \|\mathbf{z}\|$$ ■

La medida dada a una matriz conforme a la norma natural describe cómo la matriz extiende los vectores unitarios relacionados con esa norma. La extensión máxima es la norma de la matriz. Las normas matriciales que consideraremos aquí tienen las formas

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty, \qquad \text{la norma } l_\infty$$

y

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2, \qquad \text{la norma } l_2.$$

Figura 7.4

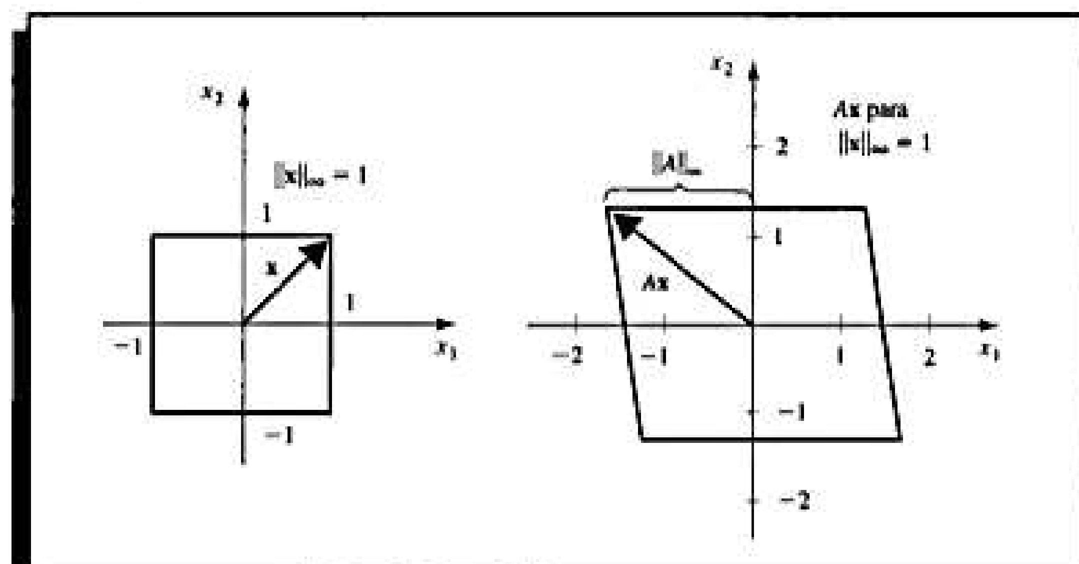
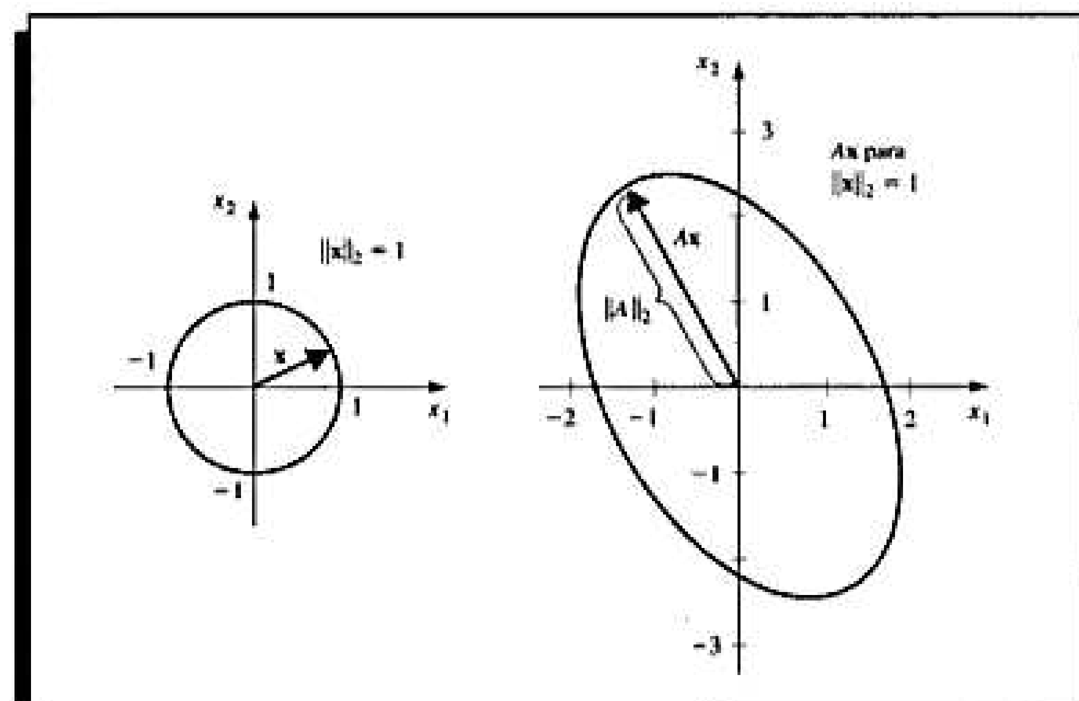


Figura 7.5



En las figuras 7.4 y 7.5 se incluye un ejemplo de estas normas cuando $n = 2$.

La norma l_∞ de una matriz tiene una representación interesante respecto a sus elementos.

Teorema 7.11 Si $A = (a_{ij})$ es una matriz de $n \times n$, entonces

$$\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|. \qquad \blacksquare$$

Demostración Primero demostremos que $\|A\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$. Sea $\mathbf{x}$ un vector columna n-dimensional con $1 = \|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i|$. Puesto que $A\mathbf{x}$ también es un vector columna n-dimensional,

$$\|A\mathbf{x}\|_\infty = \max_{1\le i\le n} |(A\mathbf{x})_i| = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \max_{1\le j\le n} |x_j|.$$

Dado que $\max_{1\le j\le n} |x_j| = \|\mathbf{x}\|_\infty = 1$, tenemos

$$\|A\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

En consecuencia,

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|. \tag{7.3}$$

Ahora necesitamos demostrar la desigualdad opuesta, $\|A\|_\infty \ge \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$. Sea p un entero tal que

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|,$$

y sea $\mathbf{x}$ el vector con las componentes

$$x_j = \begin{cases} 1, & \text{si } a_{pj} \ge 0, \\ -1, & \text{si } a_{pj} < 0. \end{cases}$$

Entonces $\|\mathbf{x}\|_\infty = 1$ y $a_{pj}x_j = |a_{pj}|$, para toda $j = 1, 2, \ldots, n$, así que

$$\|A\mathbf{x}\|_\infty = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \ge \left| \sum_{j=1}^{n} a_{pj}x_j \right| = \left| \sum_{j=1}^{n} |a_{pj}| \right| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Este resultado implica que

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty \ge \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|,$$

lo cual, junto con la desigualdad (7.3), nos da

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

■ ■ ■

EJEMPLO 5 Si

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{bmatrix},$$

entonces

$$\sum_{j=1}^{3} |a_{1j}| = |1| + |2| + |-1| = 4,$$

$$\sum_{j=1}^{3} |a_{2j}| = |0| + |3| + |-1| = 4,$$

y

$$\sum_{j=1}^{3} |a_{3j}| = |5| + |-1| + |1| = 7;$$

así que

$$\|A\|_\infty = \text{máx}\{4, 4, 7\} = 7. \qquad \blacksquare$$

En la siguiente sección veremos un método alterno con el que se obtiene la norma l_2 de una matriz.

CONJUNTO DE EJERCICIOS 7.1

1. Obtenga $\|\mathbf{x}\|_\infty$ y $\|\mathbf{x}\|_2$ para los siguientes vectores.
   a. $\mathbf{x} = (3, -4, 0, \frac{3}{2})^t$
   b. $\mathbf{x} = (2, 1, -3, 4)^t$
   c. $\mathbf{x} = (\text{sen } k, \cos k, 2^k)^t$ para un entero positivo fijo k
   d. $\mathbf{x} = (4/(k+1), 2/k^2, k^2e^{-k})^t$ para un entero positivo fijo k
2. a. Verifique que la función $\|\cdot\|_1$, definida en $\mathbb{R}^n$ por

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|,$$

   es una norma de $\mathbb{R}^n$.
   b. Obtenga $\|\mathbf{x}\|_1$ para los vectores del ejercicio 1.
   c. Demuestre que para todo $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2$.
3. Demuestre que las siguientes sucesiones son convergentes y encuentre sus límites.
   a. $\mathbf{x}^{(k)} = (1/k, e^{1-k}, -2/k^2)^t$
   b. $\mathbf{x}^{(k)} = (e^{-k}\cos k, k \text{ sen } (1/k), 3 + k^{-2})^t$
   c. $\mathbf{x}^{(k)} = (ke^{-k^2}, (\cos k)/k, \sqrt{k^2 + k} - k)^t$
   d. $\mathbf{x}^{(k)} = (e^{1/k}, (k^2+1)/(1-k^2), (1/k^2)(1 + 3 + 5 + \cdots + (2k-1)))^t$
4. Obtenga $\|\cdot\|_\infty$ para las matrices siguientes.

   a. $\begin{bmatrix} 10 & 15 \\ 0 & 1 \end{bmatrix}$ b. $\begin{bmatrix} 10 & 0 \\ 15 & 1 \end{bmatrix}$

c. $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ d. $\begin{bmatrix} 4 & -1 & 7 \\ -1 & 4 & 0 \\ -7 & 0 & 4 \end{bmatrix}$

5. Los sistemas lineales siguientes $A\mathbf{x} = \mathbf{b}$ tienen a $\mathbf{x}$ como la solución real y a $\tilde{\mathbf{x}}$ como una solución aproximada. Calcule $\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty$ y $\|A\tilde{\mathbf{x}} - \mathbf{b}\|_\infty$.

a. $\frac{1}{2}x_1 + \frac{1}{3}x_2 = \frac{1}{63}$,
$\frac{1}{3}x_1 + \frac{1}{4}x_2 = \frac{1}{168}$,
$\mathbf{x} = \left(\frac{1}{7}, -\frac{1}{6}\right)^t$,
$\tilde{\mathbf{x}} = (0.142, -0.166)^t$.

b. $x_1 + 2x_2 + 3x_3 = 1$,
$2x_1 + 3x_2 + 4x_3 = -1$,
$3x_1 + 4x_2 + 6x_3 = 2$,
$\mathbf{x} = (0, -7, 5)^t$,
$\tilde{\mathbf{x}} = (-0.33, -7.9, 5.8)^t$.

c. $x_1 + 2x_2 + 3x_3 = 1$,
$2x_1 + 3x_2 + 4x_3 = -1$,
$3x_1 + 4x_2 + 6x_3 = 2$,
$\mathbf{x} = (0, -7, 5)^t$,
$\tilde{\mathbf{x}} = (-0.2, -7.5, 5.4)^t$.

d. $0.04x_1 + 0.01x_2 - 0.01x_3 = 0.06$,
$0.2x_1 + 0.5x_2 - 0.2x_3 = 0.3$,
$x_1 + 2x_2 + 4x_3 = 11$,
$\mathbf{x} = (1.827586, 0.6551724, 1.965517)^t$,
$\tilde{\mathbf{x}} = (1.8, 0.64, 1.9)^t$.

6. La norma matricial $\|\cdot\|_1$, definida por $\|A\|_1 = \max_{\|\mathbf{x}\|_1=1} \|A\mathbf{x}\|_1$, puede calcularse mediante la fórmula

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|,$$

donde la norma vectorial $\|\cdot\|_1$ se definió en el ejercicio 2. Obtenga $\|\cdot\|_1$ para las matrices del ejercicio 4.

7. Demuestre con un ejemplo que $\|\cdot\|_{\otimes}$, definida por $\|A\|_{\otimes} = \max_{1\le i,j\le n} |a_{ij}|$, no define una norma matricial.

8. Demuestre que $\|\cdot\|_{\odot}$ definida por

$$\|A\|_{\odot} = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|,$$

es una norma matricial. Obtenga $\|\cdot\|_{\odot}$ para las matrices del ejercicio 4.

9. **a.** La norma de Frobenius (que no es una norma natural) se define para una matriz A de $n \times n$ por medio de

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}$$

Demuestre que $\|\cdot\|_F$ es una norma matricial.

b. Obtenga $\|\cdot\|_F$ para las matrices del ejercicio 4.

c. Para cualquier matriz A, demuestre que $\|A\|_2 \le \|A\|_F \le n^{1/2}\|A\|_2$.

10. En el ejercicio 9 definimos la norma de Frobenius para una matriz. Demuestre que para cualquier matriz A de $n \times n$ y para cualquier vector $\mathbf{x}$ de $\mathbb{R}^n$, $\|A\mathbf{x}\|_2 \le \|A\|_F\|\mathbf{x}\|_2$.

11. Sea S una matriz definida positiva de $n \times n$. Para toda $\mathbf{x}$ en $\mathbb{R}^n$ defina $\|\mathbf{x}\| = (\mathbf{x}^tS\mathbf{x})^{1/2}$. Demuestre que esto define una norma en $\mathbb{R}^n$. [*Sugerencia:* use la descomposición de Choleski de S para demostrar primero que $\mathbf{x}^tS\mathbf{y} = \mathbf{y}^tS\mathbf{x} \le (\mathbf{x}^tS\mathbf{x})^{1/2}(\mathbf{y}^tS\mathbf{y})^{1/2}$.]

12. Sea S una matriz real y no singular, y sea $\|\cdot\|$ cualquier norma de $\mathbb{R}^n$. Defina $\|\cdot\|'$ por medio de $\|\mathbf{x}\|' = \|S\mathbf{x}\|$. Demuestre que $\|\cdot\|'$ también es una norma en $\mathbb{R}^n$.

13. Demuestre que si $\|\cdot\|$ es una norma vectorial en $\mathbb{R}^n$, entonces $\|A\| = \text{máx}_{\|\mathbf{x}\|=1}\|A\mathbf{x}\|$ es una norma matricial.

14. El siguiente extracto de *Mathematics Magazine* [Sz] contiene una forma alterna de demostrar la desigualdad de Cauchy-Buniakowsky-Schwarz.

a. Demuestre que cuando $\mathbf{x} \ne \mathbf{0}$ y $\mathbf{y} \ne \mathbf{0}$ tenemos

$$\frac{\sum_{i=1}^n x_iy_i}{\left(\sum_{i=1}^n x_i^2\right)^{1/2}\left(\sum_{i=1}^n y_i^2\right)^{1/2}} = 1 - \frac{1}{2}\sum_{i=1}^n\left(\frac{x_i}{\left(\sum_{j=1}^n x_j^2\right)^{1/2}} - \frac{y_i}{\left(\sum_{j=1}^n y_j^2\right)^{1/2}}\right)^2.$$

b. Use el resultado de la parte (a) para demostrar que

$$\sum_{i=1}^n x_iy_i \le \left(\sum_{i=1}^n x_i^2\right)^{1/2}\left(\sum_{i=1}^n y_i^2\right)^{1/2}.$$

7.2 Vectores y valores característicos

Una matriz $n \times m$ se puede considerar como una función que usa la multiplicación de matrices para llevar vectores m-dimensionales en vectores n-dimensionales en sí mismo. En este caso, ciertas vectores no nulos $\mathbf{x}$ son paralelos a $A\mathbf{x}$, lo que significa que existe una constante λ tal que $A\mathbf{x} = \lambda\mathbf{x}$. Para estos vectores, tenemos $(A - \lambda I)\mathbf{x} = \mathbf{0}$. Existe una relación muy estrecha entre estos números λ y la posibilidad de que un método iterativo converja. En esta sección estudiaremos tal relación.

Definición 7.12 Si A es una matriz cuadrada, el polinomio definido por

$$p(\lambda) = \det(A - \lambda I)$$

recibe el nombre de **polinomio característico** de A. ■

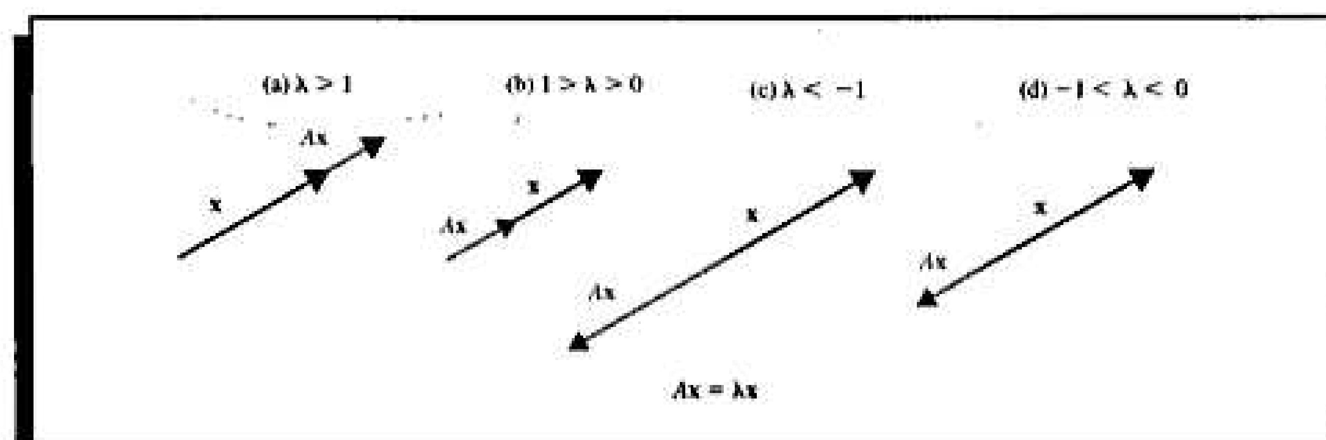
No es difícil demostrar (véase el ejercicio 7) que p es un polinomio de grado n y, en consecuencia, que tiene como máximo n ceros distintos, algunos de los cuales pueden ser complejos. Si λ es un cero de p, entonces como $\det(A - \lambda I) = 0$, el teorema 6.16 en la sección 6.4 implica que el sistema lineal definido por $(A - \lambda I)\mathbf{x} = \mathbf{0}$ tiene otra solución que no es $\mathbf{x} \ne \mathbf{0}$. Deseamos estudiar los ceros de p y las soluciones distintas de cero correspondientes a estos sistemas.

Definición 7.13 Si p es el polinomio característico de la matriz A, los ceros de p se llaman **valores característicos** o **propios** de esa matriz. Si λ es un valor característico de A y si $\mathbf{x} \ne \mathbf{0}$ tiene la

propiedad de que $(A - \lambda I)\mathbf{x} = \mathbf{0}$ entonces a $\mathbf{x}$ se le llama **vector característico** o propio de la matriz A correspondiente al valor característico λ. ■

Si $\mathbf{x}$ es un vector característico asociado al valor característico λ, entonces $A\mathbf{x} = \lambda\mathbf{x}$, por lo cual la matriz A transforma al vector $\mathbf{x}$ en un múltiplo escalar de sí misma. Si λ es real y $\lambda > 1$, entonces A tiene el efecto de extender a $\mathbf{x}$ en un factor de λ, como se observa en la figura 7.6(a). Si $0 < \lambda < 1$, entonces A reduce a $\mathbf{x}$ en un factor de λ [véase la figura 7.6(b)]. Cuando $\lambda < 0$, los efectos son semejantes [véase la figura 7.6(c) y (d)], aunque se invierte la dirección de $A\mathbf{x}$.

Figura 7.6



EJEMPLO 1 Sea

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix}.$$

Para calcular los valores característicos de A, consideremos

$$p(\lambda) = \det(A - \lambda I) = \det \begin{bmatrix} 1-\lambda & 0 & 2 \\ 0 & 1-\lambda & -1 \\ -1 & 1 & 1-\lambda \end{bmatrix} = (1-\lambda)(\lambda^2 - 2\lambda + 4).$$

Los valores característicos de A son las soluciones de $p(\lambda) = 0$, que son $\lambda_1 = 1$, $\lambda_2 = 1 + \sqrt{3}i$ y $\lambda_3 = 1 - \sqrt{3}i$.

Un vector característico $\mathbf{x}$ de A asociado a λ_1 es una solución del sistema $(A - \lambda_1 I)\mathbf{x} = \mathbf{0}$:

$$\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Por tanto,

$$2x_3 = 0, \quad -x_3 = 0 \quad \text{y} \quad -x_1 + x_2 = 0,$$

lo cual implica que

$$x_3 = 0, \quad x_2 = x_1 \quad \text{y} \quad x_1 \text{ arbitraria.}$$

La elección $x_1 = 1$ produce el vector característico $(1, 1, 0)^t$ correspondiente al valor característico $\lambda_1 = 1$. Para esta elección tenemos que $\|(1, 1, 0)^t\|_\infty = 1$.

La elección $x_1 = \sqrt{2}/2$ produce un vector característico correspondiente a λ con

$$\left\|\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0\right)^t\right\|_2 = 1.$$

Puesto que λ_2 y λ_3 son números complejos, sus vectores característicos correspondientes también lo son. A fin de encontrar un vector característico para λ_2, resolvemos el sistema

$$\begin{bmatrix} 1-(1+\sqrt{3}i) & 0 & 2 \\ 0 & 1-(1+\sqrt{3}i) & -1 \\ -1 & 1 & 1-(1+\sqrt{3}i) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Una solución de este sistema es el vector

$$\left(-\frac{2\sqrt{3}}{3}i, \frac{\sqrt{3}}{3}i, 1\right)^t.$$

De modo similar, el vector

$$\left(\frac{2\sqrt{3}}{3}i, -\frac{\sqrt{3}}{3}i, 1\right)^t.$$

es un vector característico correspondiente al valor característico $\lambda_3 = 1 - \sqrt{3}i$. ■

Maple tiene la función `Eigenvals` que calcula los valores característicos y, de manera opcional, los vectores característicos de una matriz. Para el ejemplo visto introducimos lo siguiente:

```
>with(linalg);
>A:=matrix(3,3,[1,0,2,0,1,-1,-1,1,1]);
>evalf(Eigenvals(A));
```

$$[1.000000000 + 1.732050807I, 1.000000000 - 1.732050807I, 1.000000000]$$

Con esto se calculan los valores característicos

$$\lambda_2 = 1 + \sqrt{3}i, \quad \lambda_3 = 1 - \sqrt{3}i, \quad \lambda_1 = 1.$$

Para calcular los valores y vectores característicos utilizamos

```
>evalf(Eigenvals(A,B));
```

Los valores característicos se calculan y muestran como antes; los vectores característicos se indican en las columnas de B. Si todos los valores característicos son reales, cada columna de B da un vector característico. Sin embargo, en nuestro ejemplo mostramos B aplicando

```
>evalm(B);
```

$$B = \begin{bmatrix} 1.154700538 & .6324555321\ 10^{-10} & .7453559925 \\ -.5773502680 & .1264911064\ 10^{-9} & .7453559926 \\ -.2581988896\ 10^{-19} & 1.000000000 & -.72572776\ 10^{-11} \end{bmatrix}$$

Las dos primeras columnas corresponden a las partes real e imaginaria de los vectores característicos correspondientes a los valores característicos λ_2 y λ_3. Por tanto, un vector característico de λ_2 es

$$\begin{bmatrix} 1.154700538 \\ -.5773502680 \\ -.2581988896\ 10^{-19} \end{bmatrix} + \begin{bmatrix} .6324555321\ 10^{-10} \\ .1264911064\ 10^{-9} \\ 1.000000000 \end{bmatrix} i \approx \begin{bmatrix} 1.154700538 \\ -.5773502680 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} i;$$

es decir,

$$(1.154700538, -0.5773502680, i)^t = \left(\frac{2\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, i\right)^t.$$

Dado que cualquier múltiplo de un vector característico es también un vector característico, tenemos

$$(1, -0.5, 0.8660254i)$$

como vector característico. Al multiplicar cada coordenada por $-(2\sqrt{3}/3)i$ se obtiene el vector característico del ejemplo 1:

$$\left(-\frac{2\sqrt{3}}{3}i, \frac{\sqrt{3}}{3}i, 1\right)^t.$$

Como λ_1 es real, la tercera columna de B es un vector característico correspondiente a λ_1.

Los conceptos de valores y vectores característicos se explican aquí por una razón práctica de cálculo, pero se presentan con cierta frecuencia en el estudio de los sistemas físicos. De hecho, tienen tanto interés que el capítulo 9 está dedicado íntegramente a su aproximación numérica.

Definición 7.14 El **radio espectral** $\rho(A)$ de una matriz A está definido por

$$\rho(A) = \text{máx}|\lambda|, \qquad \text{donde } \lambda \text{ es un valor característico de } A.$$

(Recuérdese que, para $\lambda = \alpha + \beta i$ complejo, tenemos $|\lambda| = (\alpha^2 + \beta^2)^{1/2}$.) ■

En el caso de la matriz del ejemplo 1,

$$\rho(A) = \text{máx}\{1, |1 + \sqrt{3}i|, |1 - \sqrt{3}i|\} = \text{máx}\{1, 2, 2\} = 2.$$

Como se puede apreciar en el siguiente teorema, el radio espectral guarda estrecha relación con la norma de una matriz.

Teorema 7.15 Si A es una matriz de $n \times n$, entonces

(i) $\|A\|_2 = [\rho(A^tA)]^{1/2}$.
(ii) $\rho(A) \le \|A\|$, para cualquier norma natural $\|\cdot\|$. ■

Demostración La demostración de la parte (i) requiere mayor información sobre los valores característicos que la que tenemos en el momento actual. Al lector que desee mayores detalles concernientes a la demostración, le recomendamos consultar [Or2, p. 21].

Para probar la parte (ii), supongamos que λ es un valor característico de A con el vector característico $\mathbf{x}$ donde $\|\mathbf{x}\| = 1$. (El ejercicio 6 nos garantiza que existe ese vector.) Dado que $A\mathbf{x} = \lambda\mathbf{x}$, para cualquier norma natural

$$|\lambda| = |\lambda| \cdot \|\mathbf{x}\| = \|\lambda\mathbf{x}\| = \|A\mathbf{x}\| \leq \|A\| \, \|\mathbf{x}\| = \|A\|.$$

Por tanto,

$$\rho(A) = \text{máx}|\lambda| \leq \|A\|.$$

■ ■ ■

La parte (i) del teorema 7.15 implica que si A es simétrica, entonces $\|A\|_2 = \rho(A)$ (véase el ejercicio 10).

Un resultado interesante y útil, que se parece a la parte (ii) del teorema 7.15, es que para cualquier matriz A y para todo $\varepsilon > 0$ existe una norma natural $\|\cdot\|$ con la propiedad de que $\rho(A) < \|A\| < \rho(A) + \varepsilon$. En consecuencia, $\rho(A)$ es la cota máxima inferior de las normas naturales de A. La demostración de este resultado viene en [Or2, p. 23].

EJEMPLO 2 Si

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix},$$

entonces

$$A^t A = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{bmatrix}.$$

Si queremos calcular $\rho(A^t A)$, necesitamos los valores característicos de A^tA. Si

$$\begin{aligned} 0 &= \det(A^t A - \lambda I) \\ &= \det \begin{bmatrix} 3-\lambda & 2 & -1 \\ 2 & 6-\lambda & 4 \\ -1 & 4 & 5-\lambda \end{bmatrix} \\ &= -\lambda^3 + 14\lambda^2 - 42\lambda = -\lambda(\lambda^2 - 14\lambda + 42), \end{aligned}$$

entonces

$$\lambda = 0 \quad \text{o} \quad \lambda = 7 \pm \sqrt{7},$$

así que

$$\|A\|_2 = \sqrt{\rho(A^t A)} = \sqrt{\text{máx}\{0, 7-\sqrt{7}, 7+\sqrt{7}\}} = \sqrt{7+\sqrt{7}} \approx 3.106.$$

■

Las operaciones del ejemplo 2 se pueden realizar con Maple, haciendo

```
>with(linalg);
>A:=matrix(3,3,[1,1,0,1,2,1,-1,1,2]);
>B:=transpose(A);
>C:=multiply(A,B);
>evalf(Eigenvals(C));
```

$$[0.1097678465\ 10^{-8}, 4.354248690, 9.645751311]$$

Como $\|A\|_2 = \sqrt{\rho(A^t A)} = \sqrt{\rho(C)}$, tenemos

$$\|A\|_2 = \sqrt{9.645751311} = 3.105760987.$$

Maple también calcula $\|A\|_2 = \sqrt{7 + \sqrt{7}}$ directamente, con el comando

```
>norm(A,2);
```

Para determinar la norma l_∞ de A, se reemplaza el último comando con

```
>norm(A,infinity);
```

Al estudiar los métodos iterativos de matrices, es muy importante saber cuándo las potencias de una matriz se vuelven pequeñas (es decir, cuándo todos los elementos se aproximan a cero). A este tipo de matrices se les llama *convergentes*.

Definición 7.16 Llamamos **convergente** a una matriz de $n \times n$ si

$$\lim_{k\to\infty} (A^k)_{ij} = 0, \qquad \text{para cada } i = 1, 2, \ldots, n \text{ y } j = 1, 2, \ldots, n. \qquad \blacksquare$$

EJEMPLO 3 Sea

$$A = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Al calcular las potencias de A obtenemos:

$$A^2 = \begin{bmatrix} \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \quad A^3 = \begin{bmatrix} \frac{1}{8} & 0 \\ \frac{3}{16} & \frac{1}{8} \end{bmatrix}, \quad A^4 = \begin{bmatrix} \frac{1}{16} & 0 \\ \frac{1}{8} & \frac{1}{16} \end{bmatrix},$$

y, en general,

$$A^k = \begin{bmatrix} (\frac{1}{2})^k & 0 \\ \frac{k}{2^{k+1}} & (\frac{1}{2})^k \end{bmatrix}.$$

Dado que

$$\lim_{k\to\infty} \left(\frac{1}{2}\right)^k = 0 \qquad \text{y} \qquad \lim_{k\to\infty} \frac{k}{2^{k+1}} = 0,$$

A es una matriz convergente. ■

En el ejemplo 3 la matriz convergente A tiene $\rho(A) = \frac{1}{2}$, porque $\frac{1}{2}$ es el único valor característico de A. Esto ejemplifica la importante conexión que existe entre el radio espectral de una matriz y su convergencia, como se verá más a fondo en el siguiente resultado.

Teorema 7.17 Las siguientes afirmaciones son equivalentes.

(i) A es una matriz convergente.

(ii) $\lim_{n\to\infty} \|A^n\| = 0$, para alguna norma natural.

(iii) $\lim_{n\to\infty} \|A^n\| = 0$, para todas las normas naturales.

(iv) $\rho(A) < 1$.

(v) $\lim_{n\to\infty} A^n \mathbf{x} = \mathbf{0}$, para toda $\mathbf{x}$. ■

La prueba de este teorema se encuentra en [IK, p. 14].

CONJUNTO DE EJERCICIOS 7.2

1. Calcule los valores característicos y los vectores característicos asociados de las siguientes matrices.

a. $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ b. $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$

c. $\begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$ d. $\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}$

e. $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ f. $\begin{bmatrix} -1 & 2 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 7 \end{bmatrix}$

g. $\begin{bmatrix} 2 & 1 & 1 \\ 2 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix}$ h. $\begin{bmatrix} 3 & 2 & -1 \\ 1 & -2 & 3 \\ 2 & 0 & 4 \end{bmatrix}$

2. Calcule el radio espectral de las matrices del ejercicio 1.
3. ¿Cuáles de las matrices del ejercicio 1 son convergentes?
4. Sean $A_1 = \begin{bmatrix} 1 & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$ y $A_2 = \begin{bmatrix} \frac{1}{2} & 0 \\ 16 & \frac{1}{2} \end{bmatrix}$. Demuestre que A_1 no es convergente, pero A_2 es convergente.
5. Obtenga $\|\cdot\|_2$ para todas las matrices del ejercicio 1.
6. Demuestre que si λ es un valor característico de una matriz A, y si $\|\cdot\|$ es una norma vectorial, entonces existe un vector característico $\mathbf{x}$ asociado a λ con $\|\mathbf{x}\| = 1$.
7. Demuestre que el polinomio característico $p(\lambda) = \det(A - \lambda I)$ para la matriz A de $n \times n$ es un polinomio de n-ésimo grado. [*Sugerencia*: desarrolle $\det(A - \lambda I)$ a lo largo del primer renglón y use inducción matemática en n.]
8. a. Demuestre que si A es una matriz de $n \times n$, entonces

$$\det A = \prod_{i=1}^{n} \lambda_i,$$

donde $\lambda_1, \ldots, \lambda_n$, son valores característicos de A. [*Sugerencia*: considere $p(0)$.]

b. Demuestre que A es singular si y sólo si $\lambda = 0$ es un valor característico de A.

9. Sea λ un valor característico de la matriz A de $n \times n$ y sea $\mathbf{x} \neq \mathbf{0}$ un vector característico asociado.

a. Demuestre que λ también es un valor característico de A^t.

b. Demuestre que para todo entero $k \geq 1$, λ^k es un valor característico de A^k con el vector característico $\mathbf{x}$.

c. Demuestre que si existe A^{-1}, entonces $1/\lambda$ es un valor característico de A^{-1} con el vector característico $\mathbf{x}$.

d. Generalice las partes (b) y (c) a $(A^{-1})^k$ para los enteros $k \geq 2$.

e. Dado el polinomio $q(x) = q_0 + q_1x + \cdots + q_kx^k$, defina $q(A)$ como la matriz $q(A) = q_0I + q_1A + \cdots + q_kA^k$. Demuestre que $q(\lambda)$ es un valor característico de $q(A)$ con el vector característico $\mathbf{x}$.

f. Sea $\alpha \neq \lambda$. Demuestre que si $A - \alpha I$ es no singular, entonces $1/(\lambda - \alpha)$ es un valor característico de $(A - \alpha I)^{-1}$ con el vector característico $\mathbf{x}$.

10. Demuestre que si A es simétrica, entonces $\|A\|_2 = \rho(A)$.

11. En el ejercicio 11 de la sección 6.3 supusimos que la contribución de un escarabajo hembra de cierto tipo a la población de escarabajos de años futuros podía expresarse mediante la matriz

$$A = \begin{bmatrix} 0 & 0 & 6 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \end{bmatrix},$$

donde el elemento del i-ésimo renglón y de la j-ésima columna representa la contribución probabilística de un escarabajo de edad j a la población de hembras de edad i en el año siguiente.

a. ¿Tiene la matriz A valores característicos reales? De ser así, determine dichos valores y los vectores característicos asociados.

b. Si una muestra de esta especie se necesitara para efectuar experimentos de laboratorio que año tras año tuvieran una proporción constante en cada grupo de edad, ¿qué criterios se impondrían a la población inicial para estar seguros de que se cumpliría este requisito?

12. Encuentre matrices A y B para las cuales $\rho(A + B) > \rho(A) + \rho(B)$. (Esto demuestra que $\rho(A)$ no puede ser una norma matricial.)

13. Demuestre que, si $\|\cdot\|$ es una norma natural cualquiera, entonces $(1/\|A^{-1}\|) \leq |\lambda| \leq \|A\|$ para todo valor característico λ de una matriz no singular A.

7.3 Métodos iterativos para resolver sistemas lineales

En esta sección describiremos los métodos iterativos de Jacobi y Gauss-Seidel, métodos clásicos que datan de fines del siglo XVIII. Los métodos iterativos rara vez se usan para resolver sistemas lineales de pequeña dimensión, ya que el tiempo necesario para conseguir una exactitud satisfactoria rebasa el que requieren los métodos directos, como el de la iluminación gaussiana. Sin embargo, en el caso de sistemas grandes con un alto porcentaje de elementos cero, son eficientes tanto en almacenamiento de computadora como en el tiempo de cómputo. Este tipo de sistemas se presentan constantemente en los análisis de circuitos y en la solución numérica de los problemas con valor en la frontera y de ecuaciones diferenciales parciales.

Un método iterativo con el cual se resuelve el sistema lineal $A\mathbf{x} = \mathbf{b}$ comienza con una aproximación inicial $\mathbf{x}^{(0)}$ a la solución $\mathbf{x}$ y genera una sucesión de vectores $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que converge a $\mathbf{x}$. Los métodos iterativos traen consigo un proceso que convierte el sistema $A\mathbf{x} = \mathbf{b}$ en otro equivalente de la forma $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ para alguna matriz fija T y un vector $\mathbf{c}$.

Luego de seleccionar el vector inicial $\mathbf{x}^{(0)}$ la sucesión de los vectores de la solución aproximada se genera calculando

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c},$$

para cada $k = 1, 2, 3, \ldots$. Este resultado debería recordarnos la iteración de punto fijo que estudiamos en el capítulo 2.

EJEMPLO 1 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned}
E_1:&\quad 10x_1 - x_2 + 2x_3 = 6,\\
E_2:&\quad -x_1 + 11x_2 - x_3 + 3x_4 = 25,\\
E_3:&\quad 2x_1 - x_2 + 10x_3 - x_4 = -11,\\
E_4:&\quad 3x_2 - x_3 + 8x_4 = 15
\end{aligned}$$

tiene la solución única $\mathbf{x} = (1, 2, -1, 1)^t$. Para convertir $A\mathbf{x} = \mathbf{b}$ en la forma $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, resolvemos la ecuación E_i para x_i con cada $i = 1, 2, 3, 4$ y así obtenemos

$$\begin{aligned}
x_1 &= \frac{1}{10}x_2 - \frac{1}{5}x_3 + \frac{3}{5},\\
x_2 &= \frac{1}{11}x_1 + \frac{1}{11}x_3 - \frac{3}{11}x_4 + \frac{25}{11},\\
x_3 &= -\frac{1}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}x_4 - \frac{11}{10},\\
x_4 &= -\frac{3}{8}x_2 + \frac{1}{8}x_3 + \frac{15}{8}.
\end{aligned}$$

Si queremos escribir esto en la forma $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, utilizamos

$$T = \begin{bmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{bmatrix} \quad \text{y} \quad \mathbf{c} = \begin{bmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{bmatrix}.$$

En el caso de una aproximación inicial, hacemos $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$. Entonces $\mathbf{x}^{(1)}$ está dado por

$$\begin{aligned}
x_1^{(1)} &= \frac{1}{10}x_2^{(0)} - \frac{1}{5}x_3^{(0)} + \frac{3}{5} = 0.6000,\\
x_2^{(1)} &= \frac{1}{11}x_1^{(0)} + \frac{1}{11}x_3^{(0)} - \frac{3}{11}x_4^{(0)} + \frac{25}{11} = 2.2727,\\
x_3^{(1)} &= -\frac{1}{5}x_1^{(0)} + \frac{1}{10}x_2^{(0)} + \frac{1}{10}x_4^{(0)} - \frac{11}{10} = -1.1000,\\
x_4^{(1)} &= -\frac{3}{8}x_2^{(0)} + \frac{1}{8}x_3^{(0)} + \frac{15}{8} = 1.8750.
\end{aligned}$$

Las iteraciones adicionales $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})^t$, se generan de manera parecida y se incluyen en la tabla 7.1.

Tabla 7.1

k	0	1	2	3	4	5	6	7	8	9	10
$x_1^{(k)}$	0.000	0.6000	1.0473	0.9326	1.0152	0.9890	1.0032	0.9981	1.0006	0.9997	1.0001
$x_2^{(k)}$	0.0000	2.2727	1.7159	2.053	1.9537	2.0114	1.9922	2.0023	1.9987	2.0004	1.9998
$x_3^{(k)}$	0.0000	−1.1000	−0.8052	−1.0493	−0.9681	−1.0103	−0.9945	−1.0020	−0.9990	−1.0004	−0.9998
$x_4^{(k)}$	0.0000	1.8750	0.8852	1.1309	0.9739	1.0214	0.9944	1.0036	0.9989	1.0006	0.9998

La decisión de parar después de diez iteraciones se basó en el criterio

$$\frac{\|\mathbf{x}^{(10)} - \mathbf{x}^{(9)}\|_\infty}{\|\mathbf{x}^{(10)}\|_\infty} = \frac{8.0 \times 10^{-4}}{1.9998} < 10^{-3}.$$

De hecho, $\|\mathbf{x}^{(10)} - \mathbf{x}\|_\infty = 0.0002$. ■

El método del ejemplo 1 se denomina **método iterativo de Jacobi**, y consiste en resolver la i-ésima ecuación en $A\mathbf{x} = \mathbf{b}$ para x_i a fin de obtener (a condición de que $a_{ii} \neq 0$)

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(-\frac{a_{ij}x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}}, \quad \text{para } i = 1, 2, \ldots, n$$

y generar cada $x_i^{(k)}$ a partir de los componentes de $\mathbf{x}^{(k-1)}$ cuando $k \geq 1$ por medio de

$$x_i^{(k)} = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \left(-a_{ij}x_j^{(k-1)} \right) + b_i}{a_{ii}}, \quad \text{para } i = 1, 2, \ldots, n. \tag{7.4}$$

El método se escribe en la forma $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ separando A en sus partes diagonales y fuera de la diagonal. Para comprobar esto, sea D la matriz diagonal cuya diagonal es la misma que A, sea $-L$ la parte estrictamente triangular inferior de la parte A y sea $-U$ la parte estrictamente triangular superior de A. Con esta notación,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

se divide en

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \cdots & -a_{n,n-1} & 0 \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

$= D - L - U.$

Entonces transformamos la ecuación $A\mathbf{x} = \mathbf{b}$, o $(D - L - U)\mathbf{x} = \mathbf{b}$, en

$$D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b}.$$

y, si D^{-1} existe, es decir, si $a_{ii} \neq 0$ para cada i, entonces

$$\mathbf{x} = D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b}.$$

Esto da origen a la forma matricial del método iterativo de Jacobi:

$$\mathbf{x}^{(k)} = D^{-1}(L + U)\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}, \quad k = 1, 2, ... \tag{7.5}$$

Al introducir la notación $T_j = D^{-1}(L + U)$ y $\mathbf{c}_j = D^{-1}\mathbf{b}$, esta técnica tiene la forma

$$\mathbf{x}^{(k)} = T_j\mathbf{x}^{(k-1)} + \mathbf{c}_j. \tag{7.6}$$

En la práctica, la ecuación (7.4) se usa en el cálculo y la (7.6) se emplea con fines teóricos.

En el algoritmo 7.1 se pone en ejecución el método iterativo de Jacobi.

ALGORITMO 7.1

Método iterativo de Jacobi

Para resolver $A\mathbf{x} = \mathbf{b}$ dada una aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones e incógnitas n; los elementos a_{ij}, $1 \leq i, j \leq n$ de la matriz A; los elementos b_i, $1 \leq i \leq n$ de $\mathbf{b}$; los elementos XO_i, $1 \leq i \leq n$ de $\mathbf{XO} = \mathbf{x}^{(0)}$; la tolerancia TOL; el número máximo de iteraciones N.

SALIDA la solución aproximada $x_1, \ldots, x_n$ o el mensaje de que se rebasó el número de iteraciones.

Paso 1 Tome $k = 1$.

Paso 2 Mientras ($k \leq N$) haga pasos 3-6.

Paso 3 Para $i = 1, \ldots, n$

$$\text{tome } x_i = \frac{-\sum_{\substack{j=1 \\ j \neq i}}^{n} (a_{ij} XO_j) + b_i}{a_{ii}}.$$

Paso 4 Si $\|\mathbf{x} - \mathbf{XO}\| < TOL$ entonces SALIDA $(x_1, \ldots, x_n)$;
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 5 Tome $k = k + 1$.

Paso 6 Para $i = 1, \ldots, n$ tome $XO_i = x_i$.

Paso 7 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

El paso 3 del algoritmo requiere que $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$. Si uno de los elementos a_{ii} es cero y si el sistema es no singular, podemos reordenar las ecuaciones de modo que ningún $a_{ii} = 0$. Si queremos acelerar la convergencia, debemos arreglar las ecuaciones de modo que a_{ii} sea lo más grande posible. Este tema se trata de manera más detallada en secciones posteriores en este capítulo.

Otro posible criterio de interrupción en el paso 4 consiste en iterar hasta que

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\|\mathbf{x}^{(k)}\|}$$

sea menor que alguna tolerancia prescrita. Para ello podemos emplear cualquier norma adecuada, siendo la usual l_∞.

Un análisis de la ecuación (7.4) sugiere una mejora del algoritmo 7.1. Si queremos calcular $x_i^{(k)}$, utilizamos las componentes de $\mathbf{x}^{(k-1)}$. Como para $i > 1$, ya se calcularon $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$, y probablemente sean mejores aproximaciones de las soluciones reales $x_1, \ldots, x_{i-1}$ que $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$, parece más razonable calcular $x_i^{(k)}$, por medio de los valores calculados más recientemente. Esto es, podemos usar

$$x_i^{(k)} = \frac{-\sum_{j=1}^{i-1} (a_{ij}x_j^{(k)}) - \sum_{j=i+1}^{n} (a_{ij}x_j^{(k-1)}) + b_i}{a_{ii}}, \tag{7.7}$$

para cada $i = 1, 2, \ldots, n$, en vez de la ecuación (7.4). A esta modificación se le llama **método iterativo de Gauss-Seidel** y se explica en el siguiente ejemplo.

EJEMPLO 2 El sistema lineal dado por

$$\begin{aligned} 10x_1 - x_2 + 2x_3 &= 6, \\ -x_1 + 11x_2 - x_3 + 3x_4 &= 25, \\ 2x_1 - x_2 + 10x_3 - x_4 &= -11, \\ 3x_2 - x_3 + 8x_4 &= 15 \end{aligned}$$

se resolvió en el ejemplo 1 por medio del método iterativo de Jacobi. Al incorporar la ecuación (7.7) en el algoritmo 7.1 se obtienen las ecuaciones empleadas con cada $k = 1, 2, \ldots,$

$$\begin{aligned} x_1^{(k)} &= \frac{1}{10}x_2^{(k-1)} - \frac{1}{5}x_3^{(k-1)} + \frac{3}{5}, \\ x_2^{(k)} &= \frac{1}{11}x_1^{(k)} + \frac{1}{11}x_3^{(k-1)} - \frac{3}{11}x_4^{(k-1)} + \frac{25}{11}, \\ x_3^{(k)} &= -\frac{1}{5}x_1^{(k)} + \frac{1}{10}x_2^{(k)} + \frac{1}{10}x_4^{(k-1)} - \frac{11}{10}, \\ x_4^{(k)} &= -\frac{3}{8}x_2^{(k)} + \frac{1}{8}x_3^{(k)} + \frac{15}{8}. \end{aligned}$$

Siendo $\mathbf{x}^{(0)} = (0, 0, 0, 0)^t$, generamos las iteraciones de la tabla 7.2.

Tabla 7.2

k	0	1	2	3	4	5
$x_1^{(k)}$	0.0000	0.6000	1.030	1.0065	1.0009	1.0001
$x_2^{(k)}$	0.0000	2.3272	2.037	2.0036	2.0003	2.0000
$x_3^{(k)}$	0.0000	−0.9873	−1.014	−1.0025	−1.0003	−1.0000
$x_4^{(k)}$	0.0000	0.8789	0.9844	0.9983	0.9999	1.0000

Dado que

$$\frac{\|\mathbf{x}^{(5)} - \mathbf{x}^{(4)}\|_\infty}{\|\mathbf{x}^{(5)}\|_\infty} = \frac{0.0008}{2.000} = 4 \times 10^{-4},$$

se acepta $\mathbf{x}^{(5)}$ como una aproximación razonable de la solución. Nótese que en el ejemplo 1 el método de Jacobi requirió el doble de iteraciones para alcanzar el mismo grado de exactitud. ■

Si queremos expresar el método de Gauss-Seidel en forma matricial, multiplicaremos ambos lados de la ecuación (7.7) por a_{ii} y reunimos todos los k-ésimos términos de iteración, lo que nos da

$$a_{i1}x_1^{(k)} + a_{i2}x_2^{(k)} + \cdots + a_{ii}x_i^{(k)} = -a_{i,i+1}x_{i+1}^{(k-1)} - \cdots - a_{in}x_n^{(k-1)} + b_i,$$

para cada $i = 1, 2, \ldots, n$. Al escribir todas las n ecuaciones obtenemos

$$\begin{aligned}
a_{11}x_1^{(k)} \qquad\qquad &= -a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)} - \cdots - a_{1n}x_n^{(k-1)} + b_1,\\
a_{21}x_1^{(k)} + -a_{22}x_2^{(k)} \qquad &= \qquad -a_{23}x_3^{(k-1)} - \cdots - a_{2n}x_n^{(k-1)} + b_2,\\
\vdots \qquad\qquad & \\
a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \cdots + a_{nn}x_n^{(k)} &= \qquad\qquad b_n;
\end{aligned}$$

y con las definiciones anteriores de D, L y U, se deduce que la forma matricial del método de Gauss-Seidel es

$$(D - L)\,\mathbf{x}^{(k)} = U\,\mathbf{x}^{(k-1)} + \mathbf{b}$$

o bien

$$\mathbf{x}^{(k)} = (D - L)^{-1}U\,\mathbf{x}^{(k-1)} + (D - L)^{-1}\,\mathbf{b}, \text{ para cada } k = 1, 2, \ldots. \tag{7.8}$$

Si usamos $T_g = (D - L)^{-1}\,U$ y $\mathbf{c}_g = (D - L)^{-1}\mathbf{b}$, el método de Gauss-Seidel tiene la forma

$$\mathbf{x}^{(k)} = T_g\mathbf{x}^{(k-1)} + \mathbf{c}_g. \tag{7.9}$$

Para que la matriz triangular inferior $D - L$ sea no singular, es necesario y suficiente que $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$.

En el algoritmo 7.2 se implanta el método de Gauss-Seidel.

ALGORITMO 7.2

Método iterativo de Gauss-Seidel

Para resolver $A\mathbf{x} = \mathbf{b}$ dada una aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones e incógnitas n; los elementos a_{ij}, $1 \leq i, j \leq n$ de la matriz A; los elementos b_i, $1 \leq i \leq n$ de $\mathbf{b}$; los elementos XO_i, $1 \leq i \leq n$ de $\mathbf{XO} = \mathbf{x}^{(0)}$; la tolerancia TOL; el número máximo de iteraciones N.

SALIDA la solución aproximada $x_1, \ldots, x_n$ o el mensaje de que se rebasó el número de iteraciones.

Paso 1 Tome $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3-6.

Paso 3 Para $i = 1, \ldots, n$

$$\text{tome } x_i = \frac{-\sum_{j=1}^{i-1} a_{ij} x_j - \sum_{j=i+1}^{n} a_{ij} XO_j + b_i}{a_{ii}}.$$

Paso 4 Si $\|\mathbf{x} - \mathbf{XO}\| < TOL$ entonces SALIDA $(x_1, \ldots, x_n)$
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 5 Tome $k = k + 1$.

Paso 6 Para $i = 1, \ldots, n$ tome $XO_i = x_i$.

Paso 7 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Los comentarios que acompañan al algoritmo 7.1 respecto a los criterios de reordenación e interrupción, también se aplican al algoritmo 7.2 de Gauss-Seidel.

Los resultados de los ejemplos 1 y 2 parecen implicar que este método es superior al de Jacobi, y generalmente es así. Hay sistemas lineales donde el método de Jacobi converge y el de Gauss-Seidel no, y hay otros donde éste converge y aquél no. (Véanse ejercicios 9 y 10.)

A fin de estudiar la convergencia de los métodos generales de iteración, consideramos la fórmula

$$\mathbf{x}^{(k)} = T\,\mathbf{x}^{(k-1)} + \mathbf{c}, \quad \text{para cada } k = 1, 2, \ldots,$$

donde $\mathbf{x}^{(0)}$ es arbitrario.

Lema 7.18 Si el radio espectral $\rho(T)$ satisface $\rho(T) < 1$, entonces existe $(I - T)^{-1}$ y

$$(I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j. \quad ■$$

Demostración Como $T\mathbf{x} = \lambda\mathbf{x}$ es verdadera exactamente cuando $(I - T)\mathbf{x} = (1 - \lambda)\mathbf{x}$ tenemos λ como un valor característico de T exactamente cuando $1 - \lambda$ es un valor característico de $I - T$. Pero $|\lambda| \leq \rho(T) < 1$ y, por tanto, $\lambda = 1$ no es un valor característico de T y 0 tampoco puede serlo de $I - T$. Por tanto, $(I - T)^{-1}$ existe.

Sea $S_m = I + T + T^2 + \cdots + T^m$. Entonces

$$(I - T)\,S_m = (1 + T + T^2 + \cdots + T^m) - (T + T^2 + \cdots + T^{m+1}) = I - T^{m+1},$$

y como T es convergente, el resultado obtenido al final de la sección 7.2 implica que

$$\lim_{m \to \infty} (I - T)S_m = \lim_{m \to \infty} (I - T^{m+1}) = I.$$

Por tanto $(I - T)^{-1} = \lim_{m \to \infty} S_m = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j$. ■ ■ ■

Teorema 7.19 Para cualquier $\mathbf{x}^{(0)} \in \mathbb{R}^n$, la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad \text{para cada } k \geq 1, \tag{7.10}$$

converge a la solución única de $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ si y sólo si $\rho(T) < 1$. ■

Demostración Supongamos primero que $\rho(T) < 1$. Entonces,

$$\begin{aligned}
\mathbf{x}^{(k)} &= T\mathbf{x}^{(k-1)} + \mathbf{c} \\
&= T(T\mathbf{x}^{(k-2)} + \mathbf{c}) + \mathbf{c} \\
&= T^2\mathbf{x}^{(k-2)} + (T + I)\mathbf{c} \\
&\vdots \\
&= T^k\mathbf{x}^{(0)} + (T^{k-1} + \cdots + T + I)\mathbf{c}.
\end{aligned}$$

Puesto que $\rho(T) < 1$, la matriz T es convergente y

$$\lim_{k\to\infty} T^k\mathbf{x}^{(0)} = \mathbf{0}.$$

El lema 7.18 implica que

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \lim_{k\to\infty} T^k\mathbf{x}^{(0)} + \left(\sum_{j=0}^{\infty} T^j\right)\mathbf{c} = \mathbf{0} + (I - T)^{-1}\mathbf{c} = (I - T)^{-1}\mathbf{c}.$$

Por tanto, la sucesión $\{\mathbf{x}^{(k)}\}$ converge al vector $\mathbf{x} = (I - T)^{-1}\mathbf{c}$ y $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

Para mostrar el recíproco, mostraremos que para cada $\mathbf{z} \in \mathbb{R}^n$, tenemos $\lim_{k\to\infty} T^k\mathbf{z} = \mathbf{0}$. Por el teorema 7.17, esto es equivalente a $\rho(T) < 1$.

Sea $\mathbf{z}$ un vector arbitrario, y $\mathbf{x}$ la única solución de $\mathbf{x} = T\mathbf{x} + \mathbf{c}$. Definimos $\mathbf{x}^{(0)} = \mathbf{x} - \mathbf{z}$, y para $k \geq 1$, $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$. Entonces $\{\mathbf{x}^{(k)}\}$ converge a $\mathbf{x}$. Además,

$$\mathbf{x} - \mathbf{x}^{(k)} = (T\mathbf{x} + \mathbf{c}) - (T\mathbf{x}^{(k-1)} + \mathbf{c}) = T(\mathbf{x} - \mathbf{x}^{(k-1)}),$$

de modo que

$$\mathbf{x} - \mathbf{x}^{(k)} = T(\mathbf{x} - \mathbf{x}^{(k-1)}) = T^2(\mathbf{x} - \mathbf{x}^{(k-2)}) = \cdots = T^k(\mathbf{x} - \mathbf{x}^{(0)}) = T^k\mathbf{z}.$$

Por tanto

$$\lim_{k\to\infty} T^k\mathbf{z} = \lim_{k\to\infty} T^k(\mathbf{x} - \mathbf{x}^{(0)}) = \lim_{k\to\infty} (\mathbf{x} - \mathbf{x}^{(k)}) = \mathbf{0}.$$

Como $\mathbf{z} \in \mathbb{R}^n$ era arbitrario, esto implica que T es una matriz convergente y que $\rho(T) < 1$.

■ ■ ■

La comprobación del siguiente corolario se parece a las demostraciones del corolario 2.4, se incluye en el ejercicio 11.

Corolario 7.20 Si $\|T\| < 1$ para toda norma matricial natural y si $\mathbf{c}$ es un vector cualquiera, entonces la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ converge, para cualquier $\mathbf{x}^{(0)} \in \mathbb{R}^n$, a un vector $\mathbf{x} \in \mathbb{R}^n$, y las siguientes cotas de error son válidas:

(i) $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \|T\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\|$;

(ii) $\|\mathbf{x} - \mathbf{x}^{(k)}\| \leq \dfrac{\|T\|^k}{1 - \|T\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$. ■

Hemos visto que los métodos iterativos de Jacobi y de Gauss-Seidel pueden escribirse como

$$\mathbf{x}^{(k)} = T_j \mathbf{x}^{(k-1)} + \mathbf{c}_j \quad \text{y} \quad \mathbf{x}^{(k)} = T_g \mathbf{x}^{(k-1)} + \mathbf{c}_g,$$

por medio de las matrices

$$T_j = D^{-1}(L + U) \quad \text{y} \quad T_g = (D - L)^{-1}U.$$

Si $\rho(T_j)$ o $\rho(T_g)$ son menores que 1, entonces la sucesión correspondiente $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ convergirá en la solución $\mathbf{x}$ de $A\mathbf{x} = \mathbf{b}$. Por ejemplo, el esquema de Jacobi tiene

$$\mathbf{x}^{(k)} = D^{-1}(L + U)\,\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b},$$

y, si $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converge en $\mathbf{x}$, entonces

$$\mathbf{x} = D^{-1}(L + U)\,\mathbf{x} + D^{-1}\mathbf{b}.$$

Lo anterior significa que

$$D\mathbf{x} = (L + U)\,\mathbf{x} + \mathbf{b} \quad \text{y} \quad (D - L - U)\,\mathbf{x} = \mathbf{b}.$$

Puesto que $D - L - U = A$, la solución $\mathbf{x}$ satisface $A\mathbf{x} = \mathbf{b}$.

Ahora podemos proporcionar condiciones de suficiencia fácilmente verificables para la convergencia de los métodos de Jacobi y de Gauss-Seidel. (Para demostrar la convergencia del esquema de Jacobi, véase el ejercicio 12, y para el de Gauss-Seidel consúltese [Or2, p. 120].)

Teorema 7.21 Si A es estrictamente diagonal dominante, entonces con cualquier elección de $\mathbf{x}^{(0)}$, tanto el método de Jacobi como el de Gauss-Seidel dan sucesiones $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ que convergen a la solución única de $A\mathbf{x} = \mathbf{b}$. ■

En el corolario 7.20 vemos la relación entre la rapidez de la convergencia y el radio espectral de la matriz de iteración T. Como las desigualdades se mantienen para toda norma matricial natural, del enunciado que viene después del teorema 7.15 se deduce que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \approx \rho(T)^k \|\mathbf{x}^{(0)} - \mathbf{x}\|. \tag{7.11}$$

Por tanto, conviene seleccionar el método iterativo con mínimo $\rho(T) < 1$ para un sistema particular $A\mathbf{x} = \mathbf{b}$. No existen resultados generales que nos digan cuál de los dos métodos, si el de Jacobi o el de Gauss-Seidel, será más eficaz en un sistema lineal arbitrario. Sin embargo, en casos especiales sí conocemos la respuesta, como se demuestra en el siguiente teorema. La demostración de este resultado viene en [Y, pp. 120-127].

Teorema 7.22 **(Stein-Rosenberg)**

Si $a_{ij} \leq 0$, para cada $i \neq j$ y si $a_{ii} > 0$ para cada $i = 1, 2, \ldots, n$, entonces será válida una y sólo una de las siguientes afirmaciones:

a. $0 \leq \rho(T_g) < \rho(T_j) < 1$;
b. $1 < \rho(T_j) < \rho(T_g)$;
c. $\rho(T_j) = \rho(T_g) = 0$;
d. $\rho(T_j) = \rho(T_g) = 1$. ■

En el caso especial que se describe en el teorema 7.22, vemos en la parte (a) que cuando un método da convergencia, entonces ambos la dan, y el método de Gauss-Seidel converge más rápidamente que el de Jacobi. La parte (b) indica que, cuando un método diverge, entonces ambos divergen, y la divergencia es más pronunciada en el de Gauss-Seidel.

La rapidez de convergencia de un procedimiento depende del radio espectral de la matriz relacionada con el método; por ello, una forma de seleccionar un procedimiento que acelere la convergencia consiste en seleccionar un método cuya matriz asociada tenga un radio espectral mínimo. Antes de describir un procedimiento para escogerlo, debemos explicar un nuevo medio de medir el grado en que una aproximación de la solución de un sistema lineal difiere de la verdadera solución. El método usa el vector que se describe en la siguiente definición.

Definición 7.23 Supongamos que $\tilde{\mathbf{x}} \in \mathbb{R}^n$ es una aproximación a la solución del sistema lineal definido por $A\mathbf{x} = \mathbf{b}$. El **vector residual** de $\tilde{\mathbf{x}}$ respecto a este sistema es $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$. ■

En procedimientos como el método de Jacobi o de Gauss-Seidel, un vector residual se asocia a cada cálculo de una componente de la aproximación al vector solución. El método tiene por objeto generar una sucesión de aproximaciones que harán que los vectores residuales asociados converjan rápidamente a cero. Supóngase que

$$\mathbf{r}_i^{(k)} = (r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)})^t$$

denota el vector residual del método de Gauss-Seidel correspondiente al vector solución $\mathbf{x}_i^{(k)}$, aproximado definido por

$$\mathbf{x}_i^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)})^t.$$

La m-ésima componente de $\mathbf{r}_i^{(k)}$ es

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)}, \qquad (7.12)$$

o, en forma equivalente,

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)},$$

para toda $m = 1, 2, \ldots, n$.
En particular, la i-ésima componente de $\mathbf{r}_i^{(k)}$ es

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - a_{ii} x_i^{(k-1)}.$$

así que

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}. \tag{7.13}$$

Pero recuerde que, en el método de Gauss-Seidel, se decide que $x_i^{(k)}$ sea

$$x_i^{(k)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right], \tag{7.14}$$

de modo que la ecuación (7.13) puede reescribirse así

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}.$$

En consecuencia, el método de Gauss-Seidel puede caracterizarse escogiendo $x_i^{(k)}$ para que satisfaga

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}. \tag{7.15}$$

Podemos deducir otra conexión entre los vectores residuales y el método de Gauss-Seidel. Consideremos el vector residual $\mathbf{r}_{i+1}^{(k)}$, asociado al vector $\mathbf{x}_{i+1}^{(k)} = (x_1^{(k)}, \ldots, x_i^{(k)}, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)})^t$. Conforme a (7.12), la i-ésima componente de $\mathbf{r}_{i+1}^{(k)}$ es

$$\begin{aligned} r_{i,i+1}^{(k)} &= b_i - \sum_{j=1}^{i} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} - a_{ii}x_i^{(k)}. \end{aligned}$$

La ecuación (7.14) implica que $r_{i,i+1}^{(k)} = 0$. Así pues, en cierto modo el método de Gauss-Seidel se caracteriza también por seleccionar $x_{i+1}^{(k)}$ de manera que la i-ésima componente $\mathbf{r}_{i+1}^{(k)}$ sea cero.

Sin embargo, el seleccionar $x_{i+1}^{(k)}$ de modo que una coordenada del vector residual sea cero no suele ser la forma más eficiente de disminuir el tamaño global del vector $\mathbf{r}_{i+1}^{(k)}$. Al modificar el procedimiento de Gauss-Seidel como se da en la ecuación (7.15)

$$x_i^{(k)} = x_i^{(k-1)} + \omega\frac{r_{ii}^{(k)}}{a_{ii}}, \tag{7.16}$$

para ciertas opciones de ω positivo la norma del vector residual se reduce, y se logra una convergencia significativamente más rápida.

A los métodos que contienen la ecuación (7.16) se les llama **métodos de relajación.** En las selecciones de ω con $0 < \omega < 1$, reciben el nombre de **métodos de subrelajación** y pueden servir para obtener la convergencia de algunos sistemas que no son convergentes con el método de Gauss-Seidel. Para las selecciones de ω con $1 < \omega$, a los procedimientos se les denomina **métodos de sobrerrelajación** y sirven para acelerar la convergencia de sistemas que son convergentes con el método de Gauss-Seidel. Estos procedimientos se designan con la abreviatura **SOR** (***Successive Over-Relaxation***, sobrerrelajación sucesiva) y son de gran utilidad en la resolución de sistemas lineales que se presentan en la solución numérica de algunas ecuaciones diferenciales parciales.

Antes de mostrar las ventajas del método SOR, es conveniente señalar que mediante la ecuación (7.13) con $m = i$, podemos reformular la ecuación (7.16) con fines de cálculo, como sigue:

$$x_i^{(k)} + (1 - \omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}\right].$$

Para determinar la matriz del método SOR, reescribimos lo anterior como

$$a_{ii}x_i^{(k)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} = (1 - \omega)\, a_{ii}x_i^{(k-1)} - \omega\sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)} + \omega b_i,$$

así que

$$(D - \omega L)\mathbf{x}^{(k)} = [(1 - \omega)D + \omega U]\mathbf{x}^{(k-1)} + \omega\mathbf{b}$$

o bien

$$\mathbf{x}^{(k)} = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]\mathbf{x}^{(k-1)} + \omega(D - \omega L)^{-1}\mathbf{b}. \quad (7.17)$$

Si utilizamos $T_\omega = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]$ y $\mathbf{c}_\omega = \omega(D - \omega L)^{-1}\mathbf{b}$, podemos expresar el método SOR en la forma

$$\mathbf{x}^{(k)} = T_\omega\mathbf{x}^{(k-1)} + \mathbf{c}_\omega. \quad (7.18)$$

EJEMPLO 3 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned} 4x_1 + 3x_2 \qquad &= 24, \\ 3x_1 + 4x_2 - x_3 &= 30, \\ -x_2 + 4x_3 &= -24, \end{aligned}$$

tiene la solución $(3, 4, -5)^t$. Los métodos de Gauss-Seidel y SOR con $\omega = 1.25$ se emplearán para resolver este sistema, usando $\mathbf{x}^{(0)} = (1, 1, 1)^t$ en ambos métodos. Las ecuaciones para el método de Gauss-Seidel son

$$\begin{aligned} x_1^{(k)} &= -0.75x_2^{(k-1)} + 6, \\ x_2^{(k)} &= -0.75x_1^{(k)} + 0.25x_3^{(k-1)} + 7.5, \\ x_3^{(k)} &= 0.25x_2^{(k)} - 6, \end{aligned}$$

y las ecuaciones para el método SOR con $\omega = 1.25$ son

$$\begin{aligned} x_1^{(k)} &= -0.25x_1^{(k-1)} - 0.9375\,x_2^{(k-1)} + 7.5, \\ x_2^{(k)} &= -0.9375x_1^{(k)} - 0.25\,x_2^{(k-1)} + 0.3125\,x_3^{(k-1)} + 9.375, \\ x_3^{(k)} &= 0.3125x_2^{(k)} - 0.25\,x_3^{(k-1)} - 7.5. \end{aligned}$$

En las tablas 7.3 y 7.4 se incluyen las primeras siete iteraciones para cada método. Para que las iteraciones tengan una exactitud de siete cifras decimales, el método de Gauss-Seidel requiere 34 iteraciones, en contraste con las 14 iteraciones que exige el método de sobrerrelajación con $\omega = 1.25$. ■

Tabla 7.3 Gauss-Seidel

k	0	1	2	3	4	5	6	7
$x_1^{(k)}$	1	5.250000	3.1406250	3.0878906	3.0549316	3.0343323	3.0214577	3.0134110
$x_2^{(k)}$	1	3.812500	3.8828125	3.9267578	3.9542236	3.9713898	3.9821186	3.9888241
$x_3^{(k)}$	1	−5.046875	−5.0292969	−5.0183105	−5.0114441	−5.0071526	−5.0044703	−5.0027940

Tabla 7.4 SOR con $\omega = 1.25$

k	0	1	2	3	4	5	6	7
$x_1^{(k)}$	1	6.312500	2.6223145	3.1333027	2.9570512	3.0037211	2.9963276	3.0000498
$x_2^{(k)}$	1	3.5195313	3.9585266	4.0102646	4.0074838	4.0029250	4.0009262	4.0002586
$x_3^{(k)}$	1	−6.6501465	−4.6004238	−5.0966863	−4.9734897	−5.0057135	−4.9982822	−5.0003486

La pregunta obvia que hemos de preguntar es cómo se selecciona el valor apropiado de ω. Aunque no se conoce una respuesta completa a esta pregunta para el sistema lineal general $n \times n$, los resultados siguientes serán útiles en algunos casos.

Teorema 7.24 **(Kahan)**

Si $a_{ii} \neq 0$ para cada $i = 1, 2, \ldots, n$, entonces $\rho(T_\omega) \geq |\omega - 1|$. Ello significa que el método SOR puede converger sólo si $0 < \omega < 2$. ■

La demostración de este teorema viene en el ejercicio 13. La demostración de los dos siguientes resultados se encuentra en [Or2, pp. 123-133]. Estos resultados se utilizarán en el capítulo 12.

Teorema 7.25 **(Ostrowski-Reich)**

Si A es una matriz definida positiva y si $0 < \omega < 2$, entonces el método SOR converge para cualquier elección del vector inicial aproximado $\mathbf{x}^{(0)}$. ■

Teorema 7.26 Si A es una matriz definida positiva y tridiagonal, entonces $\rho(T_g) = [\rho(T_j)]^2 < 1$, y la elección óptima de ω para el método SOR es

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}}.$$

Con esta elección de ω, tenemos $\rho(T_\omega) = \omega - 1$. ■

EJEMPLO 4 En el ejemplo 3, la matriz dada

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix},$$

es definida positiva y tridiagonal, de modo que se aplica el teorema 7.26. Dado que:

$$T_j = D^{-1}(L+U) = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.75 & 0 \\ -0.75 & 0 & 0.25 \\ 0 & 0.25 & 0 \end{bmatrix},$$

tenemos

$$T_j - \lambda I = \begin{bmatrix} -\lambda & -0.75 & 0 \\ -0.75 & -\lambda & 0.25 \\ 0 & 0.25 & -\lambda \end{bmatrix},$$

así que

$$\det(T_j - \lambda I) = -\lambda(\lambda^2 - 0.625).$$

Por tanto,

$$\rho(T_j) = \sqrt{0.625}$$

y

$$\omega = \frac{2}{1+\sqrt{1-[\rho(T_j)]^2}} = \frac{2}{1+\sqrt{1-0.625}} \approx 1.24.$$

Esto explica la rápida convergencia que se obtuvo en el ejemplo 1 mediante $\omega = 1.25$. ■

Concluimos esta sección con el algoritmo 7.3, con el cual se ejecuta el método SOR.

ALGORITMO 7.3

Método SOR

Para resolver $A\mathbf{x} = \mathbf{b}$ dado el parámetro ω y la aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones e incógnitas n; los elementos a_{ij}, $1 \le i, j \le n$, de la matriz A, los elementos b_i, $1 \le i \le n$, de b; los elementos XO_i, $1 \le i \le n$, de $\mathbf{XO} = \mathbf{x}^{(0)}$; el parámetro ω, la tolerancia TOL; el número máximo de iteraciones N.

SALIDA la solución aproximada $x_1, \ldots, x_n$ o el mensaje de que se rebasó el número de iteraciones.

Paso 1 Tome $k = 1$.

Paso 2 Mientras ($k \le N$) haga los pasos 3-6.

Paso 3 Para $i = 1, \ldots, n$

$$\text{tome } x_i = (1-\omega)XO_i + \frac{\omega\left(-\sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}XO_j + b_i\right)}{a_{ii}}.$$

Paso 4 Si $\|\mathbf{x} - \mathbf{XO}\| < TOL$ entonces SALIDA $(x_1, \ldots, x_n)$;
(*Procedimiento terminado exitosamente*).
PARAR.

Paso 5 Tome $k = k + 1$.

Paso 6 Para $i = 1, \ldots, n$ tome $XO_i = x_i$.

Paso 7 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

CONJUNTO DE EJERCICIOS 7.3

1. Obtenga las dos primeras iteraciones del método de Jacobi para los siguientes sistemas lineales, usando $\mathbf{x}^{(0)} = \mathbf{0}$:

a. $$\begin{aligned} 3x_1 - x_2 + x_3 &= 1, \\ 3x_1 + 6x_2 + 2x_3 &= 0, \\ 3x_1 + 3x_2 + 7x_3 &= 4. \end{aligned}$$

b. $$\begin{aligned} 10x_1 - x_2 &= 9, \\ -x_1 + 10x_2 - 2x_3 &= 7, \\ - 2x_2 + 10x_3 &= 6. \end{aligned}$$

c. $$\begin{aligned} 10x_1 + 5x_2 &= 6, \\ 5x_1 + 10x_2 - 4x_3 &= 25, \\ - 4x_2 + 8x_3 - x_4 &= -11, \\ - x_3 + 5x_4 &= -11. \end{aligned}$$

d. $$\begin{aligned} 4x_1 + x_2 - x_3 + x_4 &= -2, \\ x_1 + 4x_2 - x_3 - x_4 &= -1, \\ -x_1 - x_2 + 5x_3 + x_4 &= 0, \\ x_1 - x_2 + x_3 + 3x_4 &= 1. \end{aligned}$$

e. $$\begin{aligned} 4x_1 + x_2 + x_3 + x_5 &= 6, \\ -x_1 - 3x_2 + x_3 + x_4 &= 6, \\ 2x_1 + x_2 + 5x_3 - x_4 - x_5 &= 6, \\ -x_1 - x_2 - x_3 + 4x_4 &= 6, \\ 2x_2 - x_3 + x_4 + 4x_5 &= 6. \end{aligned}$$

f. $$\begin{aligned} 4x_1 - x_2 - x_4 &= 0, \\ -x_1 + 4x_2 - x_3 - x_5 &= 5, \\ - x_2 + 4x_3 - x_6 &= 0, \\ -x_1 + 4x_4 - x_5 &= 6, \\ - x_2 - x_4 + 4x_5 - x_6 &= -2, \\ - x_3 - x_5 + 4x_6 &= 6. \end{aligned}$$

2. Repita el ejercicio 1 empleando el método de Gauss-Seidel.
3. Aplique el método de Jacobi para resolver los sistemas lineales del ejercicio 1, con $TOL = 10^{-3}$ en la norma l_∞.
4. Repita el ejercicio 3 utilizando el algoritmo de Gauss-Seidel.
5. Obtenga las dos primeras iteraciones del método SOR con $\omega = 1.1$ para los siguientes sistemas lineales usando $\mathbf{x}^{(0)} = \mathbf{0}$:

a. $$\begin{aligned} 3x_1 - x_2 + x_3 &= 1, \\ 3x_1 + 6x_2 + 2x_3 &= 0, \\ 3x_1 + 3x_2 + 7x_3 &= 4. \end{aligned}$$

b. $$\begin{aligned} 10x_1 - x_2 &= 9, \\ -x_1 + 10x_2 - 2x_3 &= 7, \\ - 2x_2 + 10x_3 &= 6. \end{aligned}$$

c. $$\begin{aligned} 10x_1 + 5x_2 &= 6, \\ 5x_1 + 10x_2 - 4x_3 &= 25, \\ -4x_2 + 8x_3 - x_4 &= -11, \\ -x_3 + 5x_4 &= -11. \end{aligned}$$

d. $$\begin{aligned} 4x_1 + x_2 - x_3 + x_4 &= -2, \\ x_1 + 4x_2 - x_3 - x_4 &= -1, \\ -x_1 - x_2 + 5x_3 + x_4 &= 0, \\ x_1 - x_2 + x_3 + 3x_4 &= 1. \end{aligned}$$

e. $$\begin{aligned} 4x_1 + x_2 + x_3 + x_5 &= 6, \\ -x_1 - 3x_2 + x_3 + x_4 &= 6, \\ 2x_1 + x_2 + 5x_3 - x_4 - x_5 &= 6, \\ -x_1 - x_2 - x_3 + 4x_4 &= 6, \\ 2x_2 - x_3 + x_4 + 4x_5 &= 6. \end{aligned}$$

f. $$\begin{aligned} 4x_1 - x_2 - x_4 &= 0, \\ -x_1 + 4x_2 - x_3 - x_5 &= 5, \\ -x_2 + 4x_3 - x_6 &= 0, \\ -x_1 + 4x_4 - x_5 &= 6, \\ -x_2 - x_4 + 4x_5 - x_6 &= -2, \\ -x_3 - x_5 + 4x_6 &= 6. \end{aligned}$$

6. Repita el ejercicio 1 usando $\omega = 1.3$.

7. Aplique el método SOR con $\omega = 1.2$ para resolver los sistemas lineales del ejercicio 5 con una tolerancia $TOL = 10^{-3}$ en la norma l_∞.

8. Determine cuáles matrices del ejercicio 5 son tridiagonales y definidas positivas. Repita el ejercicio 7 con estas matrices utilizando la selección óptima de ω.

9. El sistema lineal

$$\begin{aligned} 2x_1 - x_2 + x_3 &= -1, \\ 2x_1 + 2x_2 + 2x_3 &= 4, \\ -x_1 - x_2 + 2x_3 &= -5 \end{aligned}$$

tiene la solución $(1, 2, -1)^t$.

a. Muestre que $\rho(T_j) = \frac{\sqrt{5}}{2} > 1$.

b. Muestre que el método de Jacobi con $\mathbf{x}^{(0)} = \mathbf{0}$ no da una buena aproximación después de 25 iteraciones.

c. Muestre que $\rho(T_g) = \frac{1}{2}$.

d. Use el método de Gauss-Seidel con $\mathbf{x}^{(0)} = \mathbf{0}$p la solución al sistema lineal hasta 10^{-5} en la norma l_∞.

10. El sistema lineal

$$\begin{aligned} x_1 + 2x_2 - 2x_3 &= 7, \\ x_1 + x_2 + x_3 &= 2, \\ 2x_1 + 2x_2 + x_3 &= 5 \end{aligned}$$

tiene la solución $(1, 2, -1)^t$.

a. Muestre que $\rho(T_j) = 0$.

b. Use el método de Jacobi con $\mathbf{x}^{(0)} = \mathbf{0}$ para aproximar la solución al sistema lineal hasta 10^{-5} en la norma l_∞.

c. Muestre que $\rho(T_g) = 2$.

d. Muestre que el método de Gauss-Seidel aplicado como en la parte (b) no da una buena aproximación después de 25 iteraciones.

11. **a.** Demuestre que

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \leq \|T\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\| \quad \text{y} \quad \|\mathbf{x}^{(k)} - \mathbf{x}\| \leq \frac{\|T\|^k}{1 - \|T\|} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|,$$

donde T es una matriz de $n \times n$ con $\|T\| < 1$ y

$$\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}, \quad k = 1, 2, \ldots,$$

Con $\mathbf{x}^{(0)}$ arbitrario, $\mathbf{c} \in \mathbb{R}^n$ y $\mathbf{x} = T\mathbf{x} + \mathbf{c}$.

b. Aplique, cuando sea posible, las cotas al ejercicio 1 usando la norma l_∞.

12. Demuestre que si A es estrictamente diagonal dominante, entonces $\|T\|_\infty < 1$.

13. Demuestre el teorema 7.24. [*Sugerencia*: si $\lambda_1, \ldots, \lambda_n$ son valores característicos de T_ω, entonces $\det T_\omega = \Pi_{i=1}^n \lambda_i$. Dado que $\det D^{-1} = \det(D - \omega L)^{-1}$ y el determinante de un producto de matrices es el producto de los determinantes de los factores, el resultado se deduce de la ecuación (7.17).]

14. Suponga que un objeto puede estar en cualquiera de los $n + 1$ puntos uniformemente espaciados $x_0, x_1, \ldots, x_n$ en la línea. Cuando un objeto se encuentra ubicado en el lugar x_i tendrá las mismas probabilidades de desplazarse hacia x_{i-1} o hacia x_{i+1} y no puede dirigirse directamente hacia ningún otro lugar. Considere las probabilidades $\{P_i\}_{i=0}^n$ de que un objeto que parte del lugar x_i llegue al extremo izquierdo x_0 antes de alcanzar el extremo derecho x_n. Por supuesto, $P_0 = 1$ y $P_n = 0$. Dado que el objeto puede desplazarse hacia x_i sólo a partir de x_{i-1} o de x_{i+1} y lo hace con la probabilidad $\frac{1}{2}$ para cada uno de esos lugares,

$$P_i = \frac{1}{2} P_{i-1} + \frac{1}{2} P_{i+1}, \qquad \text{para cada } i = 1, 2, \ldots, n-1.$$

a. Demuestre que

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 & \cdots & \cdots & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} & \ddots & & \vdots \\ 0 & -\frac{1}{2} & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & \cdots & \cdots & 0 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_{n-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

b. Resuelva este sistema usando $n = 10$, 50 y 100.

c. Cambie las probabilidades a α y $1 - \alpha$ para el movimiento hacia la izquierda y derecha, respectivamente, y después derive el sistema lineal semejante al de la parte (a).

d. Repita la parte (b) con $\alpha = \frac{1}{3}$.

15. Utilice todos los métodos aplicables de esta sección para obtener las soluciones del sistema lineal de $A\mathbf{x} = \mathbf{b}$ con una exactitud de 10^{-5} en la norma l_∞.

$$a_{i,j} = \begin{cases} 2i, & \text{cuando } j = i \text{ e } i = 1, 2, \ldots, 80, \\ 0.5i, & \text{cuando } \begin{cases} j = i + 2 \text{ e } i = 1, 2, \ldots, 78, \\ j = i - 2 \text{ e } i = 3, 4, \ldots, 80, \end{cases} \\ 0.25i, & \text{cuando } \begin{cases} j = i + 4 \text{ e } i = 1, 2, \ldots, 76, \\ j = i - 4 \text{ e } i = 5, 6, \ldots, 80, \end{cases} \\ 0, & \text{de lo contrario,} \end{cases}$$

y aquellos de $\mathbf{b}$ son $b_i = \pi$, para cada $i = 1, 2, \ldots, 80$.

16. Suponga que A es una matriz definida positiva.

a. Demuestre que podemos escribir $A = D - L - L^t$, donde D es diagonal con $d_{ii} > 0$ para cada $1 \leq i \leq n$ y donde L es triangular. Más aún, demuestre que $D - L$ es no singular.

b. Sean $T_g = (D - L)^{-1}L^t$ y $P = A - T_g^t A T_g$. Demuestre que P es simétrica.

c. Demuestre que T_g también puede escribirse como $T_g = I - (D - L)^{-1}A$.

d. Sea $Q = (D - L)^{-1}A$. Demuestre que $T_g = I - Q$ y que $P = Q^t[AQ^{-1} - A + (Q^t)^{-1}A]Q$.

e. Demuestre que $P = Q^t D Q$ y que P es una matriz definida positiva.

f. Sea λ un valor característico de T_g con el vector característico $\mathbf{x} \neq \mathbf{0}$. Use la parte (b) para demostrar que $\mathbf{x}^t P \mathbf{x} > 0$ implica que $|\lambda| < 1$.

g. Demuestre que T_g es convergente y que el método de Gauss-Seidel converge.

17. Extienda el método de demostración en el ejercicio 14 al método SOR con $0 < \omega < 2$.

18. Las fuerzas que actúan sobre la estructura de un puente, descritas al inicio de este capítulo, satisfacen las ecuaciones de la siguiente tabla:

Junta	Componentes horizontales	Componentes verticales
①	$-F_1 + \frac{\sqrt{2}}{2}f_1 + f_2 = 0$	$\frac{\sqrt{2}}{2}f_1 + F_2 = 0$
②	$-\frac{\sqrt{2}}{2}f_1 + \frac{\sqrt{3}}{2}f_4 = 0$	$-\frac{\sqrt{2}}{2}f_1 - f_3 + \frac{1}{2}f_4 = 0$
③	$-f_2 + f_5 = 0$	$f_3 - 10\,000 = 0$
④	$-\frac{\sqrt{3}}{2}f_4 - f_5 = 0$	$\frac{1}{2}f_4 - F_3 = 0$

Este sistema lineal puede expresarse en forma matricial.

$$\begin{bmatrix} -1 & 0 & 0 & \frac{\sqrt{2}}{2} & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & \frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & 0 & -1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\sqrt{3}}{2} & -1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 10\,000 \\ 0 \\ 0 \end{bmatrix}.$$

a. Explique por qué se reordenó el sistema de ecuaciones.

b. Aproxime la solución del sistema lineal resultante con una exactitud de 10^{-2} en la norma l_∞, utilizando como aproximación inicial el vector cuyos elementos son todos unos con (i) el método de Gauss-Seidel, (ii) el método de Jacobi y (iii) el método SOR con $\omega = 1.25$.

7.4 Estimaciones de error y refinamiento iterativo

Desde el punto de vista intuitivo parece razonable que, si $\tilde{\mathbf{x}}$ es una aproximación a la solución $\mathbf{x}$ de $A\mathbf{x} = \mathbf{b}$ y si el vector residual $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ tiene la propiedad de que $\|\mathbf{r}\| = \|\mathbf{b} - A\tilde{\mathbf{x}}\|$ es pequeño, entonces $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ también será pequeño. A menudo este es el caso, pero algunos sistemas, que en la práctica se presentan con frecuencia, no poseen esta propiedad.

EJEMPLO 1 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

tiene la solución única $\mathbf{x} = (1, 1)^t$. La aproximación deficiente $\tilde{\mathbf{x}} = (3, 0)^t$ tiene el vector residual

$$\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.0002 \end{bmatrix},$$

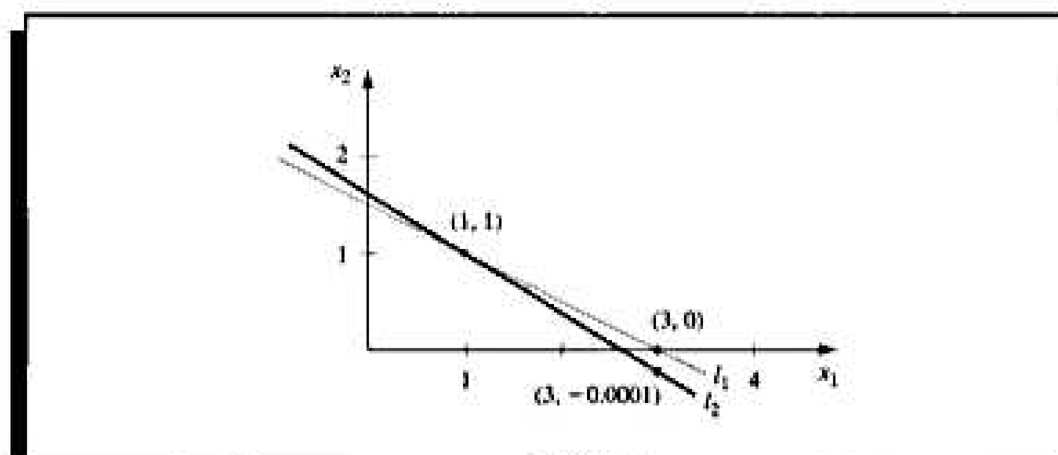
de modo que $\|\mathbf{r}\|_\infty = 0.0002$. Aunque la norma del vector residual es pequeña, la aproximación $\tilde{\mathbf{x}} = (3, 0)^t$ es evidentemente muy deficiente; de hecho, $\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty = 2$. ■

La dificultad del ejemplo 1 se explica muy fácilmente haciendo notar que la solución del sistema representa la intersección de las líneas

$$l_1: \quad x_1 + 2x_2 = 3 \quad \text{y} \quad l_2: \quad 1.0001\,x_1 + 2\,x_2 = 3.0001.$$

El punto (3, 0) se encuentra en l_1, y las líneas son casi paralelas. Esto significa que (3, 0) también está cerca de l_2, a pesar de que difiere significativamente del punto de intersección (1, 1). (Véase la Fig. 7.7.)

Figura 7.7



Evidentemente el ejemplo 1 se elaboró para demostrar las dificultades que pueden surgir y que de hecho surgen. De no haber sido las líneas casi coincidentes, cabría esperar que un vector residual pequeño implicase una aproximación exacta.

En el contexto general, no podemos recurrir a la geometría del sistema para obtener un indicio de cuándo pueden presentarse problemas. Sin embargo, sí podemos conseguir esta información si consideramos las normas de la matriz A y su inversa.

Teorema 7.27 Supongamos que $\tilde{\mathbf{x}}$ es una aproximación a la solución de $A\mathbf{x} = \mathbf{b}$, que A es una matriz no singular y que $\mathbf{r}$ es el vector residual de $\tilde{\mathbf{x}}$. Entonces, para toda norma natural,

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \|\mathbf{r}\| \cdot \|A^{-1}\|$$

y si $\mathbf{x} \neq \mathbf{0}$ y $\mathbf{b} \neq \mathbf{0}$,

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \tag{7.19}$$

■

Demostración Dado que $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = A\mathbf{x} - A\tilde{\mathbf{x}}$ y A es no singular, $\mathbf{x} - \tilde{\mathbf{x}} = A^{-1}\mathbf{r}$, el Teorema 7.11 de la sección 7.1 implica que

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \leq \|A^{-1}\| \cdot \|\mathbf{r}\|.$$

Más aún, como $\mathbf{b} = A\mathbf{x}$, tenemos $\|\mathbf{b}\| \leq \|A\| \cdot \|\mathbf{x}\|$ y, por tanto, $1/\|\mathbf{x}\| \leq \|A\|/\|\mathbf{b}\|$ y

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\|A\| \cdot \|A^{-1}\|}{\|\mathbf{b}\|} \|\mathbf{r}\|.$$

■ ■ ■

Las desigualdades del teorema 7.27 implican que las cantidades $\|A^{-1}\|$ y $\|A\| \cdot \|A^{-1}\|$ ofrecen un indicio de la conexión entre el vector residual y la exactitud de la aproximación. En general, el error relativo $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$ es de gran interés y, de acuerdo con la desigualdad (7.19), está acotado por el producto de $\|A\| \cdot \|A^{-1}\|$ con el residual relativo de esta aproximación $\|\mathbf{r}\|/\|\mathbf{b}\|$. En esta aproximación puede usarse cualquier norma adecuada; el único requisito es que se utilice de manera uniforme en todo el proceso.

Definición 7.28 El **número de condición** de una matriz no singular A relativo a la norma $\|\cdot\|$ es

$$K(A) = \|A\| \cdot \|A^{-1}\|.$$

■

Con esta notación, las desigualdades del teorema 7.27 se convierten en

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq K(A) \frac{\|\mathbf{r}\|}{\|A\|}$$

y en

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq K(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

Para toda matriz no singular A y norma natural $\|\cdot\|$,

$$1 = \|I\| = \|A \cdot A^{-1}\| \leq \|A\| \cdot \|A^{-1}\| = K(A).$$

Una matriz A es **bien condicionada** si $K(A)$ está cerca a 1 y es **mal condicionada** si $K(A)$ es significativamente mayor que 1. Dentro de este contexto, el término condición se refiere a la seguridad relativa de que un vector residual pequeño implica una solución aproximada exacta correspondiente.

EJEMPLO 2 La matriz del sistema considerado en el ejemplo 1 fue

$$A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}.$$

que tiene $\|A\|_\infty = 3.0001$. Esta norma no se considerará grande. Sin embargo,

$$A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix}, \quad \text{así } \|A^{-1}\|_\infty = 20000,$$

y, para la norma infinito, $K(A) = (20000)(3.0001) = 60002$. En este ejemplo, el tamaño del número de condición nos impedirá tomar decisiones apresuradas sobre la precisión, basadas en el residuo de la aproximación. ■

En Maple el número de condición K_∞ puede calcularse así:

```
>with(linalg) ;
>A:=matrix(2,2,[1,2,1.0001,2]);
>cond(A);
```

60002.00000

Aunque el número de condición de una matriz depende por completo de las normas de la matriz y de su inversa, en la práctica el cálculo de la inversa está sujeto al error de redondeo y depende de la exactitud con que se efectúen los cálculos. Si las operaciones incluyen la aritmética con t dígitos significativos de precisión, el número aproximado de condición de la matriz A es la norma de la matriz multiplicada por la norma de la aproximación a la inversa de A, la cual se obtiene empleando la aritmética con t dígitos. De hecho, este número de condición también depende del método con que se calcule la inversa de A.

Si suponemos que la solución aproximada del sistema lineal $A\mathbf{x} = \mathbf{b}$ se determina empleando la aritmética con t dígitos y la eliminación gaussiana, podemos demostrar (véase [FM, pp. 45-47) que el vector residual $\mathbf{r}$ para la aproximación $\tilde{\mathbf{x}}$ tiene

$$\|\mathbf{r}\| \approx 10^{-t}\,\|A\| \cdot \|\tilde{\mathbf{x}}\|. \tag{7.20}$$

A partir de la aproximación anterior puede obtenerse una estimación del número efectivo de condición en la aritmética de t dígitos, sin invertir la matriz A. En la práctica, la aproximación supone que todas las operaciones realizadas en la eliminación gaussiana se efectúan con la aritmética de t dígitos, pero que las que se requieren para determinar el residuo se realizan con la aritmética de doble precisión (es decir, de $2t$ dígitos). Con esta técnica no se incrementa mucho el número de operaciones de cálculo; además se elimina considerablemente la pérdida de exactitud que implica la resta de números casi iguales que ocurre en el cálculo del residuo.

La aproximación del número de condición con t dígitos $K(A)$ proviene de la consideración del sistema lineal

$$A\mathbf{y} = \mathbf{r}.$$

La solución de este sistema se puede aproximar fácilmente, pues ya se calcularon los multiplicadores de la eliminación gaussiana. De hecho $\tilde{\mathbf{y}}$, la solución aproximada de $A\mathbf{y} = \mathbf{r}$, satisface

$$\tilde{\mathbf{y}} \approx A^{-1}\mathbf{r} = A^{-1}(\mathbf{b} - A\tilde{\mathbf{x}}) = A^{-1}\mathbf{b} - A^{-1}A\tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}; \tag{7.21}$$

y

$$\mathbf{x} \approx \tilde{\mathbf{x}} + \tilde{\mathbf{y}}.$$

Por tanto, $\tilde{\mathbf{y}}$ es una estimación del error producido cuando $\tilde{\mathbf{x}}$ aproxima la solución $\mathbf{x}$ al sistema original. Las ecuaciones (7.20) y (7.21) implican que

$$\|\tilde{\mathbf{y}}\| \approx \|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| \leq \|A^{-1}\| \cdot \|\mathbf{r}\| \approx \|A^{-1}\| \, (10^{-t}\|A\| \cdot \|\tilde{\mathbf{x}}\|) = 10^{-t}\, \|\tilde{\mathbf{x}}\| \, K(A).$$

Esto nos da una aproximación del número de condición que interviene en la solución del sistema $A\mathbf{x} = \mathbf{b}$ usando la eliminación gaussiana y el tipo de aritmética de dígito t que acabamos de describir

$$K(A) \approx \frac{\|\tilde{\mathbf{y}}\|}{\|\tilde{\mathbf{x}}\|} 10^t. \tag{7.22}$$

EJEMPLO 3 El sistema lineal dado por

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$$

tiene la solución exacta $\mathbf{x} = (1, 1, 1)^t$.

Con la eliminación gaussiana y con la aritmética de redondeo a cinco dígitos podemos obtener las matrices aumentadas

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 & \vdots & 15913 \\ 0 & -10596 & 16.501 & \vdots & 10580 \\ 0 & -7451.4 & 6.5250 & \vdots & -7444.9 \end{bmatrix}$$

y

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 & \vdots & 15913 \\ 0 & -10596 & 16.501 & \vdots & -10580 \\ 0 & 0 & -5.0790 & \vdots & -4.7000 \end{bmatrix}.$$

La solución aproximada de este sistema es

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t.$$

El vector residual correspondiente a $\tilde{\mathbf{x}}$ se calcula con doble precisión y resulta ser

$$\begin{aligned} \mathbf{r} &= \mathbf{b} - A\tilde{\mathbf{x}} \\ &= \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix} - \begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} 1.20001 \\ 0.99991 \\ 0.92538 \end{bmatrix} \\ &= \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix} - \begin{bmatrix} 15913.00518 \\ 28.26987086 \\ 8.611560367 \end{bmatrix} = \begin{bmatrix} -0.00518 \\ 0.27412914 \\ -0.186160367 \end{bmatrix}, \end{aligned}$$

por tanto,

$$\|\mathbf{r}\|_\infty = 0.27413.$$

La estimación del número de condición dado en la explicación anterior se obtiene resolviendo primero el sistema $A\mathbf{y} = \mathbf{r}$ para $\tilde{\mathbf{y}}$:

$$\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -0.00518 \\ 0.27413 \\ -0.18616 \end{bmatrix}.$$

Lo anterior significa que $\tilde{\mathbf{y}} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t$. Al utilizar la estimación en la ecuación (7.22), obtenemos

$$K(A) \approx 10^5 \frac{\|\tilde{\mathbf{y}}\|_\infty}{\|\tilde{\mathbf{x}}\|_\infty} = \frac{10^5(0.20008)}{1.2001} = 16672. \tag{7.23}$$

Si queremos determinar el número de condición *exacto* de A, primero debemos construir A^{-1}. La aritmética de redondeo a cinco dígitos aplicada a los cálculos produce la aproximación:

$$A^{-1} = \begin{bmatrix} -1.1701 \times 10^{-4} & -1.4983 \times 10^{-1} & 8.5416 \times 10^{-1} \\ 6.2782 \times 10^{-5} & 1.2124 \times 10^{-4} & -3.0662 \times 10^{-4} \\ -8.6631 \times 10^{-5} & 1.3846 \times 10^{-1} & -1.9689 \times 10^{-1} \end{bmatrix}.$$

El teorema 7.11 implica que $\|A^{-1}\|_\infty = 1.0041$ y $\|A\|_\infty = 15934$.

En consecuencia, la matriz A mal condicionada tiene

$$K(A) = (1.0041)(15934) = 15999.$$

La estimación en (7.23) está muy cerca de $K(A)$ y requiere un esfuerzo de cálculo mucho menor.

Puesto que se conoce la solución real $\mathbf{x} = (1, 1, 1)^t$ para este sistema, podemos calcular ambas cosas.

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty = 0.2001 \quad \text{y} \quad \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{0.2001}{1} = 0.2001.$$

Las cotas de error del teorema 7.27 para estos valores son

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty \le K(A) \frac{\|\mathbf{r}\|_\infty}{\|A\|_\infty} = \frac{(15999)(0.27413)}{15934} = 0.27525$$

y

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \le K(A) \frac{\|\mathbf{r}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{(15999)(0.27413)}{15913} = 0.27561. \quad \blacksquare$$

En la ecuación (7.21) utilizamos la estimación $\tilde{\mathbf{y}} \approx \mathbf{x} - \tilde{\mathbf{x}}$, donde $\tilde{\mathbf{y}}$ es la solución aproximada del sistema $A\mathbf{y} = \mathbf{r}$. En general, $\tilde{\mathbf{x}} + \tilde{\mathbf{y}}$ es una aproximación más exacta a la solución del sistema lineal $A\mathbf{x} = \mathbf{b}$ que la aproximación original $\tilde{\mathbf{x}}$. El método en que se aplica esta suposición recibe el nombre de **refinamiento iterativo**, o mejora iterativa, y consiste en efectuar las iteraciones en el sistema cuyo lado derecho es el vector residual de las aproximaciones sucesivas hasta conseguir resultados con una exactitud satisfactoria.

El proceso se aplica mediante la aritmética de t dígitos y, si $K_\infty(A) \approx 10^q$, entonces después de k iteraciones de refinamiento iterativo, la solución tiene aproximadamente el más pequeño de los t y $k(t - q)$ dígitos correctos. Si el sistema está bien condicionado, una o dos iteraciones indicarán que la solución es exacta. Los sistemas mal condicionados también pueden mejorarse mucho, salvo en el caso de que la matriz A esté tan mal condicionada con $K_\infty(A) > 10^t$. En tal caso deberá emplearse el aumento de precisión en los cálculos.

En el algoritmo 7.4 se pone en ejecución el método de refinamiento iterativo.

ALGORITMO 7.4

Refinamiento iterativo

Para aproximar la solución del sistema lineal $A\mathbf{x} = \mathbf{b}$:

ENTRADA el número de ecuaciones e incógnitas n; los elementos a_{ij}, $1 \leq i, j \leq n$ de la matriz A; los elementos b_i, $1 \leq i \leq n$ de $\mathbf{b}$; el número máximo de iteraciones N; la tolerancia TOL; el número de dígitos de precisión t.

SALIDA la aproximación $\mathbf{xx} = (xx_i, \ldots, xx_n)^t$ o el mensaje de que se rebasó el número de iteraciones y una aproximación $COND$ a $K_\infty(A)$.

Paso 0 Resuelva el sistema $A\mathbf{x} = \mathbf{b}$ para $x_1, \ldots, x_n$ por la eliminación gaussiana guardando los multiplicadores m_{ji}, $j = i + 1, i + 2, \ldots, n$, $i = 1, 2, \ldots, n - 1$ y señalando intercambios de renglones.

Paso 1 Tome $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3-9.

Paso 3 Para $i = 1, 2, \ldots, n$ (Calcule $\mathbf{r}$.)

tome $r_i = b_i - \sum_{j=1}^{n} a_{ij}x_j$.

(*Realice los cálculos con aritmética de doble precisión.*)

Paso 4 Resuelva el sistema lineal $A\mathbf{y} = \mathbf{r}$ usando la eliminación gaussiana en el mismo orden que en el paso 0.

Paso 5 Para $i = 1, \ldots, n$ tome $xx_i = x_i + y_i$.

Paso 6 Si $k = 1$ entonces tome $COND = \dfrac{\|\mathbf{y}\|_\infty}{\|\mathbf{xx}\|_\infty} 10^t$.

Paso 7 Si $\|\mathbf{x} - \mathbf{xx}\|_\infty < TOL$ entonces SALIDA ($\mathbf{xx}$);
SALIDA ($COND$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 8 Tome $k = k + 1$.

Paso 9 Para $i = 1, \ldots, n$ tome $x_i = xx_i$.

Paso 10 SALIDA ('Número máximo de iteraciones excedido');
SALIDA ($COND$);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Si se emplea la aritmética de t dígitos, un procedimiento adecuado para interrumpir el procedimiento en el paso 7 consiste en iterar hasta que $|y_i^{(k)}| \leq 10^{-t}$, para cada $i = 1, 2, \ldots, n$.

EJEMPLO 4 En el ejemplo 3 encontramos la aproximación del problema en cuestión mediante la aritmética de cinco dígitos y la eliminación gaussiana es

$$\tilde{\mathbf{x}}^{(1)} = (1.2001, 0.99991, 0.92538)^t$$

y la solución de $A\mathbf{y} = \mathbf{r}^{(1)}$ es

$$\tilde{\mathbf{y}}^{(1)} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t.$$

De acuerdo con el paso 5 de este algoritmo,

$$\tilde{\mathbf{x}}^{(2)} = \tilde{\mathbf{x}}^{(1)} + \tilde{\mathbf{y}}^{(1)} = (1.0000, 1.0000, 0.99999)^t,$$

y el error real de esta aproximación es

$$\|\mathbf{x} - \tilde{\mathbf{x}}^{(2)}\|_\infty = 1 \times 10^{-5}.$$

Al aplicar la técnica sugerida para interrumpir el algoritmo, calculamos $\mathbf{r}^{(2)} = \mathbf{b} - A\tilde{\mathbf{x}}^{(2)}$ y resolvemos el sistema $A\mathbf{y}^{(2)} = \mathbf{r}^{(2)}$, lo cual nos da

$$\tilde{\mathbf{y}}^{(2)} = (1.5002 \times 10^{-9}, 2.0951 \times 10^{-10}, 1.0000 \times 10^{-5})^t.$$

Y como $\|\mathbf{y}^{-(2)}\|_\infty \leq 10^{-5}$, llegamos a la conclusión de que

$$\tilde{\mathbf{x}}^{(3)} = \tilde{\mathbf{x}}^{(2)} + \tilde{\mathbf{y}}^{(2)} = (10.000, 1.0000, 1.0000)^t$$

es suficientemente exacto, lo cual sin duda es correcto. ■

A lo largo de esta sección hemos supuesto que en el sistema lineal $A\mathbf{x} = \mathbf{b}$, la matriz A y el vector $\mathbf{b}$ pueden representarse de manera exacta. Desde un punto de vista realista, los elementos a_{ij} y b_j se modificarán o perturbarán en una cantidad δa_{ij} y δb_j, lo cual hará que se resuelva el sistema lineal

$$(A + \delta A)\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$$

en vez de $A\mathbf{x} = \mathbf{b}$. En condiciones normales, si $\|\delta A\|$ y $\|\delta\mathbf{b}\|$ son pequeñas (en el orden de 10^{-t}), la aritmética de t dígitos debería producir una solución $\tilde{\mathbf{x}}$ para la cual $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ es correspondientemente pequeña. Sin embargo, en el caso de sistemas mal condicionados hemos visto que aunque A y $\mathbf{b}$ pueden representarse con exactitud, los errores de redondeo pueden hacer que $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ sea grande. El siguiente teorema relaciona las perturbaciones del sistema lineal con el número de condición de una matriz. La demostración de este resultado se encuentra en [Or2, p. 33].

Teorema 7.29 Supongamos que A es una matriz no singular y que

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}.$$

La solución $\tilde{\mathbf{x}}$ de $(A + \delta A)\,\tilde{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ aproxima la solución $\mathbf{x}$ de $A\mathbf{x} = \mathbf{b}$ con una estimación de error

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{K(A)\|A\|}{\|A\| - K(A)\,\|\delta A\|}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta A\|}{\|A\|}\right). \tag{7.24}$$

■

La estimación en la desigualdad (7.24) establece que, si la matriz A está bien condicionada (es decir, si $K(A)$ no es demasiado grande), entonces cambios pequeños de A y $\mathbf{b}$ producirán cambios pequeños en la solución $\mathbf{x}$. En cambio, si A está mal condicionada, entonces los cambios pequeños de A y $\mathbf{b}$ pueden producir también cambios grandes en $\mathbf{x}$.

El teorema es independiente del procedimiento numérico usado para resolver $A\mathbf{x} = \mathbf{b}$. Mediante el análisis de error hacia atrás de Wilkinson (véase [Wil1] o [Wil2]) podemos demostrar que, si la eliminación gaussiana con pivoteo se emplea para resolver $A\mathbf{x} = \mathbf{b}$ en la aritmética de t dígitos, la solución numérica $\tilde{\mathbf{x}}$ es la solución real de un sistema lineal:

$$(A + \delta A)\,\tilde{\mathbf{x}} = \mathbf{b}, \quad \text{donde} \quad \|\delta A\|_\infty \le f(n)10^{1-t} \max_{i,j,k} |a_{i,j}^{(k)}|.$$

Wilkinson descubrió que en la práctica $f(n) \approx n$ y, en el peor de los casos, $f(n) \le 1.01(n^3 + 3n^2)$.

CONJUNTO DE EJERCICIOS 7.4

1. Calcule los números de condición de las siguientes matrices en relación con $\|\cdot\|_\infty$.

a. $\begin{bmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix}$

b. $\begin{bmatrix} 3.9 & 1.6 \\ 6.8 & 2.9 \end{bmatrix}$

c. $\begin{bmatrix} 1 & 2 \\ 1.00001 & 2 \end{bmatrix}$

d. $\begin{bmatrix} 1.003 & 58.09 \\ 5.550 & 321.8 \end{bmatrix}$

e. $\begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix}$

f. $\begin{bmatrix} 0.04 & 0.01 & -0.01 \\ 0.2 & 0.5 & -0.2 \\ 1 & 2 & 4 \end{bmatrix}$

2. Los siguientes sistemas lineales $A\mathbf{x} = \mathbf{b}$ tienen a $\mathbf{x}$ como solución real y a $\tilde{\mathbf{x}}$ como solución aproximada. Con los resultados del ejercicio 1 calcule $\|\mathbf{x} - \tilde{\mathbf{x}}\|_\infty$ y $K_\infty(A)\,\frac{\|\mathbf{b} - A\tilde{\mathbf{x}}\|_\infty}{\|A\|_\infty}$.

a. $\frac{1}{2}x_1 + \frac{1}{3}x_2 = \frac{1}{63}$,
$\frac{1}{3}x_1 + \frac{1}{4}x_2 = \frac{1}{168}$,
$\mathbf{x} = \left(\frac{1}{7}, -\frac{1}{6}\right)^t$,
$\tilde{\mathbf{x}} = (0.142, -0.166)^t$.

b. $3.9x_1 + 1.6x_2 = 5.5$,
$6.8x_1 + 2.9x_2 = 9.7$,
$\mathbf{x} = (1, 1)^t$,
$\tilde{\mathbf{x}} = (0.98, 1.1)^t$.

c. $x_1 + 2x_2 = 3$,
$1.0001x_1 + 2x_2 = 3.0001$,
$\mathbf{x} = (1, 1)^t$,
$\tilde{\mathbf{x}} = (0.96, 1.02)^t$.

d. $1.003x_1 + 58.09x_2 = 68.12$,
$5.550x_1 + 321.8x_2 = 377.3$,
$\mathbf{x} = (10, 1)^t$,
$\tilde{\mathbf{x}} = (-10, 1)^t$.

e. $x_1 - x_2 - x_3 = 2\pi,$
$x_2 - x_3 = 0,$
$-x_3 = \pi.$
$\mathbf{x} = (0, -\pi, -\pi)^t,$
$\tilde{\mathbf{x}} = (-0.1, -3.15, -3.14)^t.$

f. $0.04x_1 + 0.01x_2 - 0.01x_3 = 0.06,$
$0.2x_1 + 0.5x_2 - 0.2x_3 = 0.3$
$x_1 + 2x_2 + 4x_3 = 11,$
$\mathbf{x} = (1.827586, 0.6551724, 1.965517)^t,$
$\tilde{\mathbf{x}} = (1.8, 0.64, 1.9)^t.$

3. El sistema lineal

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

tiene la solución $(1, 1)^t$. Transforme A ligeramente en

$$\begin{bmatrix} 1 & 2 \\ 0.9999 & 2 \end{bmatrix},$$

y considere el sistema lineal

$$\begin{bmatrix} 1 & 2 \\ 0.9999 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}.$$

Calcule la solución nueva mediante la aritmética de redondeo a cinco dígitos y compare después el error real con el estimado (7.24). ¿Es A una matriz mal condicionada?

4. El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{bmatrix} 1 & 2 \\ 1.00001 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.00001 \end{bmatrix}$$

tiene la solución $(1, 1)^t$. Use la aritmética de redondeo a siete dígitos para obtener la solución del sistema perturbado

$$\begin{bmatrix} 1 & 2 \\ 1.000011 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.00001 \\ 3.00003 \end{bmatrix}$$

y después compare el error real con el estimado (7.24). ¿Es A una matriz mal condicionada?

5. a. En una computadora, use la aritmética de precisión simple para resolver el siguiente sistema lineal por medio de la eliminación gaussiana con el algoritmo de sustitución hacia atrás 6.1.

$$\begin{aligned} \frac{1}{3}x_1 - \frac{1}{3}x_2 - \frac{1}{3}x_3 - \frac{1}{3}x_4 - \frac{1}{3}x_5 &= 1 \\ \frac{1}{3}x_2 - \frac{1}{3}x_3 - \frac{1}{3}x_4 - \frac{1}{3}x_5 &= 0 \\ \frac{1}{3}x_3 - \frac{1}{3}x_4 - \frac{1}{3}x_5 &= -1 \\ \frac{1}{3}x_4 - \frac{1}{3}x_5 &= 2 \\ \frac{1}{3}x_5 &= 7 \end{aligned}$$

b. Calcule el número de condición de la matriz para el sistema en relación con $\|\cdot\|_\infty$.

c. Encuentre la solución exacta del sistema lineal.

6. La matriz de Hilbert de $n \times n$, $H^{(n)}$, definida por

$$H_{ij}^{(n)} = \frac{1}{i + j - 1}, \quad 1 \le i, j \le n,$$

es una matriz mal condicionada que se presenta al resolver las ecuaciones normales de los coeficientes del polinomio de mínimos cuadrados (véase el ejemplo 1 de la sección 8.2).

a. Demuestre que

$$[H^{(4)}]^{-1} = \begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix},$$

y calcule $K_\infty(H^{(4)})$.

b. Demuestre que

$$[H^{(5)}]^{-1} = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix},$$

y calcule $K_\infty(H^{(5)})$.

c. Resuelva el sistema lineal

$$H^{(4)} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

con la aritmética de redondeo a cinco dígitos y después compare el error real con el estimado en (7.24).

7. Demuestre que si B es singular, entonces

$$\frac{1}{K(A)} \le \frac{\|A - B\|}{\|A\|}.$$

[*Sugerencia*: existe un vector con $\|\mathbf{x}\| = 1$, tal que $B\mathbf{x} = \mathbf{0}$. Derive la estimación utilizando $\|A\mathbf{x}\| \ge \|\mathbf{x}\| / \|A^{-1}\|$.]

8. Use el ejercicio 7 para estimar los números de condición de las siguientes matrices:

a. $\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}$ **b.** $\begin{bmatrix} 3.9 & 1.6 \\ 6.8 & 2.9 \end{bmatrix}$

9. Use la aritmética de redondeo a cuatro dígitos para calcular la inversa H^{-1} de la matriz H de Hilbert de 3×3, y luego calcule $\hat{H} = (H^{-1})^{-1}$. Determine $\|H - \hat{H}\|_\infty$.

7.5 El método del gradiente conjugado

El método del gradiente conjugado de Hestenes y Stiefel [HS] fue desarrollado originalmente como un método directo diseñado para revolver un sistema lineal $n \times n$ definido positivo. Como método directo, por lo general es inferior a la eliminación gaussiana con pivoteo, pues ambos métodos requieren n pasos para determinar una solución y los pasos del método del gradiente conjugado tienen un costo mayor en cálculos que los de la eliminación gaussiana.

Sin embargo, el método del gradiente conjugado es muy útil como método iterativo de aproximación para resolver sistemas esparcidos de gran tamaño con entradas no nulas que aparecen con patrones predecibles. Estos problemas surgen con frecuencia al resolver problemas con valores en la frontera. Cuando la matriz está precondicionada para que los cálculos sean más eficaces, se obtienen buenos resultados aproximadamente en $\sqrt{n}$ pasos. Empleado de esta manera, este método es preferible sobre la eliminación gaussiana y los métodos iterativos ya analizados.

En esta sección supondremos que la matriz A es definida positiva. Usaremos la notación de *producto interior*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t\mathbf{y}, \tag{7.25}$$

donde $\mathbf{x}$ y $\mathbf{y}$ son vectores n-dimensionales. También necesitaremos algunos resultados usuales del álgebra lineal. En la sección 9.1 aparece un repaso de este material.

El siguiente resultado es una consecuencia sencilla de las propiedades de las transpuestas (véase el ejercicio 12).

Teorema 7.30 Para cualesquiera vectores $\mathbf{x}$, $\mathbf{y}$, y $\mathbf{z}$ y cualquier número real α, tenemos

(i) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$;

(ii) $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \alpha\mathbf{y} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle$;

(iii) $\langle \mathbf{x} + \mathbf{z}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle$;

(iv) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$;

(v) $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ si y sólo si $\mathbf{x} = \mathbf{0}$. ■

Cuando A es definida positiva, $\langle \mathbf{x}, A\mathbf{x} \rangle = \mathbf{x}^t A\mathbf{x} > 0$ a menos que $\mathbf{x} = \mathbf{0}$. Además, como A es simétrica, tenemos que $\mathbf{x}^t A\mathbf{y} = \mathbf{x}^t A^t \mathbf{y} = (A\mathbf{x})^t\mathbf{y}$, por lo que en este caso tenemos que para cada $\mathbf{x}$ y $\mathbf{y}$,

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \langle A\mathbf{x}, \mathbf{y} \rangle. \tag{7.26}$$

El siguiente resultado es una herramienta básica en el desarrollo del método del gradiente conjugado.

Teorema 7.31 El vector $\mathbf{x}^*$ es una solución del sistema lineal definido positivo $A\mathbf{x} = \mathbf{b}$ si y sólo si $\mathbf{x}^*$ minimiza

$$g(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle. \quad ■$$

Demostración Sean $\mathbf{x}$ y $\mathbf{v} \neq \mathbf{0}$ vectores fijos y t un número real variable. Tenemos que

$$\begin{aligned} g(\mathbf{x} + t\mathbf{v}) &= \langle \mathbf{x} + t\mathbf{v}, A\mathbf{x} + tA\mathbf{v}\rangle - 2\langle \mathbf{x} + t\mathbf{v}, \mathbf{b}\rangle \\ &= \langle \mathbf{x}, A\mathbf{x}\rangle + t\langle \mathbf{v}, A\mathbf{x}\rangle + t\langle \mathbf{x}, A\mathbf{v}\rangle, + t^2\langle \mathbf{v}, A\mathbf{v}\rangle - 2\langle \mathbf{x}, \mathbf{b}\rangle - 2t\langle \mathbf{v}, \mathbf{b}\rangle \\ &= \langle \mathbf{x}, A\mathbf{x}\rangle - 2\langle \mathbf{x}, \mathbf{b}\rangle + 2t\langle \mathbf{v}, A\mathbf{x}\rangle - 2t\langle \mathbf{v}, \mathbf{b}\rangle + t^2\langle \mathbf{v}, A\mathbf{v}\rangle, \end{aligned}$$

de modo que

$$g(\mathbf{x} + t\mathbf{v}) = g(\mathbf{x}) + 2t\langle \mathbf{v}, A\mathbf{x} - \mathbf{b}\rangle + t^2\langle \mathbf{v}, A\mathbf{v}\rangle. \tag{7.27}$$

Como $\mathbf{x}$ y $\mathbf{v}$ están fijos, definimos la función cuadrática h en t como

$$h(t) = g(\mathbf{x} + t\mathbf{v}).$$

Entonces h asume un valor mínimo cuando $h'(t) = 0$, pues su coeficiente en t^2, $\langle \mathbf{v}, A\mathbf{v}\rangle$, es positivo. Como

$$h'(t) = 2\langle \mathbf{v}, A\mathbf{x} - \mathbf{b}\rangle + 2t\langle \mathbf{v}, A\mathbf{v}\rangle,$$

el mínimo aparece cuando

$$\hat{t} = -\frac{\langle \mathbf{v}, A\mathbf{x} - \mathbf{b}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle} = \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle},$$

y de la ecuación (7.27)

$$\begin{aligned} h(\hat{t}) &= g(\mathbf{x}) - 2\frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle}\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle + \left(\frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle}\right)^2 \langle \mathbf{v}, A\mathbf{v}\rangle \\ &= g(\mathbf{x}) - \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle^2}{\langle \mathbf{v}, A\mathbf{v}\rangle}. \end{aligned}$$

Así, para cualquier vector $\mathbf{v} \neq \mathbf{0}$, tenemos que $g(\mathbf{x} + \hat{t}\mathbf{v}) < g(\mathbf{x})$ a menos que $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle = 0$, en cuyo caso $g(\mathbf{x}) = g(\mathbf{x} + \hat{t}\mathbf{v})$. Este es el resultado básico necesario para demostrar el teorema 7.31.

Supongamos que $\mathbf{x}^*$ satisface $A\mathbf{x}^* = \mathbf{b}$. Entonces $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}^*\rangle = 0$ para cualquier vector $\mathbf{v}$ y $g(\mathbf{x})$ no puede ser menor que $g(\mathbf{x}^*)$. Así, $\mathbf{x}^*$ minimiza a g.

Por otro lado, supongamos que $\mathbf{x}^*$ es un vector que minimiza g. Entonces, para cualquier vector $\mathbf{v}$, tenemos $g(\mathbf{x}^* + \hat{t}\mathbf{v}) \geq g(\mathbf{x}^*)$. Así, $\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}^*\rangle = 0$. Esto implica que $\mathbf{b} - A\mathbf{x}^* = \mathbf{0}$ y, en consecuencia, $A\mathbf{x}^* = \mathbf{b}$. ■ ■ ■

Para iniciar con el método del gradiente conjugado, elegimos $\mathbf{x}$, una solución aproximada de $A\mathbf{x}^* = \mathbf{b}$ y $\mathbf{v} \neq \mathbf{0}$, para obtener una *dirección de búsqueda* en la cual alejarse de $\mathbf{x}$ para mejorar la aproximación. Sea $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ el vector residual asociado a $\mathbf{x}$, con

$$t = \frac{\langle \mathbf{v}, \mathbf{b} - A\mathbf{x}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle} = \frac{\langle \mathbf{v}, \mathbf{r}\rangle}{\langle \mathbf{v}, A\mathbf{v}\rangle}.$$

Si $\mathbf{r} \neq \mathbf{0}$ y si $\mathbf{v}$ y $\mathbf{r}$ no son ortogonales, entonces $\mathbf{x} + t\mathbf{v}$ da un menor valor de g que $g(\mathbf{x})$ y posiblemente esté más cerca de $\mathbf{x}^*$ que $\mathbf{x}$. Esto sugiere el siguiente método.

Sea $\mathbf{x}^{(0)}$ una aproximación inicial a $\mathbf{x}^*$ y sea $\mathbf{v}^{(1)} \neq \mathbf{0}$ una dirección de búsqueda inicial. Para $k = 1, 2, 3, \ldots,$ calculamos:

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle},$$

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$$

y elegimos una nueva dirección de búsqueda $\mathbf{v}^{(k+1)}$. El objetivo es hacer esta elección de modo que la sucesión de aproximaciones $\{\mathbf{x}^{(k)}\}$ converja rápidamente a $\mathbf{x}^*$.

Para elegir las direcciones de búsqueda, vemos a g como una función de las componentes de $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$. Así,

$$g(x_1, x_2, \ldots, x_n) = \langle \mathbf{x}, A\mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{b} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - 2 \sum_{i=1}^{n} x_i b_i.$$

Al calcular las derivadas parciales con respecto de las variables componente x_k tenemos

$$\frac{\partial g}{\partial x_k}(\mathbf{x}) = 2 \sum_{i=1}^{n} a_{ki} x_i - 2 b_k.$$

Por tanto, el gradiente de g es

$$\nabla g(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}(\mathbf{x}), \frac{\partial g}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial g}{\partial x_n}(\mathbf{x}) \right)^t = 2(A\mathbf{x} - \mathbf{b}) = -2\mathbf{r},$$

donde el vector $\mathbf{r}$ es el vector residual para $\mathbf{x}$.

El cálculo de varias variables nos dice que la dirección de máximo descenso en el valor de $g(\mathbf{x})$ es la dirección dada por $-\nabla g(\mathbf{x})$; es decir, en la dirección del residual $\mathbf{r}$. El método que elige

$$\mathbf{v}^{(k+1)} = \mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$$

se llama el *método de máximo descenso*. Aunque en la sección 10.4 veremos que este método tiene su mérito para sistemas no lineales y problemas de optimización, no se usa para los sistemas lineales debido a su convergencia lenta.

Un método alternativo usa un conjunto de vectores de dirección no nulos $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ que satisfacen

$$\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0, \qquad \text{si } i \neq j.$$

Esto se llama **condición de A-ortogonalidad** y el conjunto de vectores $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ es **A-ortogonal**. No es difícil mostrar que un conjunto de vectores A-ortogonales asociados con la matriz definida positiva A es linealmente independiente. (Véase el Ejercicio 13(a).) Este conjunto de direcciones de búsqueda da

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{v}^{(k)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$

y $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}$.

El siguiente teorema muestra que esta elección de direcciones de búsqueda da la convergencia en a lo más n pasos, de modo que como método directo produce la solución exacta, suponiendo que la aritmética sea exacta.

Teorema 7.32 Sea $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ un conjunto A-ortogonal de vectores no nulos asociados con la matriz A definida positiva y sea $\mathbf{x}^{(0)}$ arbitrario. Sean

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} \qquad \text{y} \qquad \mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)},$$

para $k = 1, 2, \ldots, n$. Entonces, suponiendo una aritmética exacta, $A\mathbf{x}^{(n)} = \mathbf{b}$. ■

Demostración Como para cada $k = 1, 2, \ldots, n$,

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)},$$

tenemos

$$\begin{aligned} A\mathbf{x}^{(n)} &= A\mathbf{x}^{(n-1)} + t_n A\mathbf{v}^{(n)} \\ &= (A\mathbf{x}^{(n-2)} + t_{n-1} A\mathbf{v}^{(n-1)}) + t_n A\mathbf{x}^{(n)} \\ &\;\;\vdots \\ &= A\mathbf{x}^{(0)} + t_1 A\mathbf{v}^{(1)} + t_2 A\mathbf{v}^{(2)} + \cdots + t_n A\mathbf{v}^{(n)}. \end{aligned}$$

Al restar $\mathbf{b}$ de este resultado tenemos

$$A\mathbf{x}^{(n)} - \mathbf{b} = A\mathbf{x}^{(0)} - \mathbf{b} + t_1 A\mathbf{v}^{(1)} + t_2 A\mathbf{v}^{(2)} + \cdots + t_n A\mathbf{v}^{(n)}.$$

Ahora calculamos el producto interior de ambos lados con el vector $\mathbf{v}^{(k)}$ y usamos las propiedades del producto interior y el hecho de que A es simétrica para obtener

$$\begin{aligned} \langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_1 \langle A\mathbf{v}^{(1)}, \mathbf{v}^{(k)} \rangle + \cdots + t_n \langle A\mathbf{v}^{(n)}, \mathbf{v}^{(k)} \rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_1 \langle \mathbf{v}^{(1)}, A\mathbf{v}^{(k)} \rangle + \cdots + t_n \langle \mathbf{v}^{(n)}, A\mathbf{v}^{(k)} \rangle. \end{aligned}$$

La propiedad de A-ortogonalidad implica, para cada k,

$$\langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle = \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)} \rangle + t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle. \tag{7.28}$$

Sin embargo,

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle},$$

de modo que

$$\begin{aligned} & t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle \\ &\quad = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(k-1)} \rangle \\ &\quad = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} + A\mathbf{x}^{(0)} - A\mathbf{x}^{(1)} + \cdots - A\mathbf{x}^{(k-2)} + A\mathbf{x}^{(k-2)} - A\mathbf{x}^{(k-1)} \rangle \\ &\quad = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)} \rangle + \langle \mathbf{v}^{(k)}, A\mathbf{x}^{(0)} - A\mathbf{x}^{(1)} \rangle + \cdots + \langle \mathbf{v}^{(k)}, A\mathbf{x}^{(k-2)} - A\mathbf{x}^{(k-1)} \rangle. \end{aligned}$$

Pero para cualquier i,

$$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + t_i\mathbf{v}^{(i)} \qquad \text{y} \qquad A\mathbf{x}^{(i)} = A\mathbf{x}^{(i-1)}\, t_i A\mathbf{v}^{(i)},$$

de modo que

$$A\mathbf{x}^{(i-1)} - A\mathbf{x}^{(i)} = -\, t_i A\mathbf{v}^{(i)}.$$

Así,

$$t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle - t_1\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(1)}\rangle - \cdots - t_{k-1}\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k-1)}\rangle.$$

Debido a la A-ortogonalidad $\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(i)}\rangle = 0$ para $i \neq k$, de modo que

$$\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)}\rangle t_k = \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle.$$

De la ecuación (7.28),

$$\begin{aligned}\langle A\mathbf{x}^{(n)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle + \langle \mathbf{v}^{(k)}, \mathbf{b} - A\mathbf{x}^{(0)}\rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle + \langle \mathbf{b} - A\mathbf{x}^{(0)}, \mathbf{v}^{(k)}\rangle \\ &= \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle - \langle A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{v}^{(k)}\rangle \\ &= 0.\end{aligned}$$

El vector $A\mathbf{x}^{(n)} - \mathbf{b}$ es ortogonal al conjunto A-ortogonal de vectores $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$. Esto implica (véase el ejercicio 13(b)) que $A\mathbf{x}^{(n)} - \mathbf{b} = \mathbf{0}$. ■ ■ ■

EJEMPLO 1 Considere la matiz definida positiva

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}.$$

Sea $\mathbf{v}^{(1)} = (1, 0, 0)^t$, $\mathbf{v}^{(2)} = (-3/4, 1, 0)^t$ y $\mathbf{v}^{(3)} = (-3/7, 4/7, 1)^t$. Por cálculo directo,

$$\langle \mathbf{v}^{(1)}, A\mathbf{v}^{(2)}\rangle = \mathbf{v}^{(1)t} A\mathbf{v}^{(2)} = (1, 0, 0)\begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}\begin{bmatrix} -\frac{3}{4} \\ 1 \\ 0 \end{bmatrix} = 0,$$

$$\langle \mathbf{v}^{(1)}, A\mathbf{v}^{(3)}\rangle = (1, 0, 0)\begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}\begin{bmatrix} -\frac{3}{7} \\ \frac{4}{7} \\ 1 \end{bmatrix} = 0,$$

y

$$\langle \mathbf{v}^{(2)}, A\mathbf{v}^{(3)}\rangle = \left(-\frac{3}{4}, 1, 0\right)\begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}\begin{bmatrix} -\frac{3}{7} \\ \frac{4}{7} \\ 1 \end{bmatrix} = 0.$$

Así, $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}$ es un conjunto A-ortogonal.

El sistema lineal

$$\begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 24 \\ 30 \\ -24 \end{bmatrix},$$

tiene la solución exacta $\mathbf{x}^* = (3, 4, -5)^t$. Para aproximar esta solución, sea $\mathbf{x}^{(0)} = (0, 0, 0)^t$. Como $\mathbf{b} = (24, 30, -24)^t$, tenemos

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{b} = (24, 30, -24)^t,$$

de modo que

$$\langle \mathbf{v}^{(1)}, \mathbf{r}^{(0)} \rangle = \mathbf{v}^{(1)t}\mathbf{r}^{(0)} = 24, \quad \langle \mathbf{v}^{(1)}, A\mathbf{v}^{(1)} \rangle = 4 \qquad \text{y} \qquad t_0 = \frac{24}{4} = 6.$$

Así,

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t_0\mathbf{v}^{(1)} = (0, 0, 0)^t + 6(1, 0, 0)^t = (6, 0, 0)^t.$$

Continuando de esta manera,

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = (0, 12, -24)^t; \qquad t_1 = \frac{\langle \mathbf{v}^{(2)}, \mathbf{r}^{(1)} \rangle}{\langle \mathbf{v}^{(2)}, A\mathbf{v}^{(2)} \rangle} = \frac{12}{7/4} = \frac{48}{7};$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + t_1\mathbf{v}^{(2)} = (6, 0, 0)^t + \frac{48}{7}\left(-\frac{3}{4}, 1, 0\right)^t = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t;$$

$$\mathbf{r}^{(2)} = \mathbf{b} - A\mathbf{x}^{(2)} = \left(0, 0, -\frac{120}{7}\right); \qquad t_2 = \frac{\langle \mathbf{v}^{(3)}, \mathbf{r}^{(2)} \rangle}{\langle \mathbf{v}^{(3)}, A\mathbf{v}^{(3)} \rangle} = \frac{-120/7}{24/7} = -5;$$

y

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} + t_2\mathbf{v}^{(3)} = \left(\frac{6}{7}, \frac{48}{7}, 0\right)^t + (-5)\left(-\frac{3}{7}, \frac{4}{7}, 1\right)^t = (3, 4, -5)^t.$$

Como aplicamos la técnica $n = 3$ veces, ésta es la solución real. ■

Antes de analizar la forma de determinar el conjunto A-ortogonal, continuaremos nuestro desarrollo. El uso de un conjunto A-ortogonal $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ de vectores de dirección da lo que se llama un método de *dirección conjugada*. El siguiente teorema muestra la ortogonalidad de los vectores residuales $\mathbf{r}^{(k)}$ y los vectores de dirección $\mathbf{v}^{(j)}$. En el ejercicio 14 se considera una demostración de ese resultado mediante inducción matemática.

Teorema 7.33 Los vectores residuales $\mathbf{r}^{(k)}$, donde $k = 1, 2, \ldots, n$, para un método de dirección conjugada, satisfacen las ecuaciones

$$\langle \mathbf{r}^{(k)}, \mathbf{v}^{(j)} \rangle = 0, \qquad \text{para cada } j = 1, 2, \ldots, k.$$ ■

El método del gradiente conjugado de Hestenes y Stiefel elige las direcciones de búsqueda $\{\mathbf{v}^{(k)}\}$ durante el proceso iterativo de modo que los vectores residuales $\{\mathbf{r}^{(k)}\}$ sean mutuamente ortogonales. Para construir los vectores de dirección $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots\}$ y las aproximaciones $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$, partimos de una aproximación inicial $\mathbf{x}^{(0)}$ y usamos la dirección de máximo descenso $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$ como la primera dirección de búsqueda $\mathbf{v}^{(1)}$.

Suponga que hemos calculado las direcciones conjugadas $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k-1)}$ y las aproximaciones $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k-1)}$ con

$$\mathbf{x}^{(k-1)} = \mathbf{x}^{(k-2)} + t_{k-1}\mathbf{v}^{(k-1)},$$

donde

$$\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0 \quad \text{y} \quad \langle \mathbf{r}^{(i)}, \mathbf{r}^{(j)} \rangle = 0 \quad \text{para} \quad i \neq j.$$

Si $\mathbf{x}^{(k-1)}$ es la solución de $A\mathbf{x} = \mathbf{b}$, hemos terminado. En caso contrario $\mathbf{r}^{(k-1)} = \mathbf{b} - A\mathbf{x}^{(k-1)} \neq \mathbf{0}$ y el teorema 7.33 implica que $\langle \mathbf{r}^{(k-1)}, \mathbf{v}^{(i)} \rangle = 0$ para $i = 1, 2, \ldots, k-1$. Entonces usamos $\mathbf{r}^{(k-1)}$ para generar $\mathbf{v}^{(k)}$ haciendo

$$\mathbf{v}^{(k)} = \mathbf{r}^{(k-1)} + s_{k-1}\mathbf{v}^{(k-1)}.$$

Queremos elegir s_{k-1} de modo que

$$\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = 0.$$

Como

$$A\mathbf{v}^{(k)} = A\mathbf{r}^{(k-1)} + s_{k-1}A\mathbf{v}^{(k-1)}$$

y

$$\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = \langle \mathbf{v}^{(k-1)}, A\mathbf{r}^{(k-1)} \rangle + s_{k-1}\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k-1)} \rangle,$$

tendremos $\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k)} \rangle = 0$ cuando

$$s_{k-1} = -\frac{\langle \mathbf{v}^{(k-1)}, A\mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k-1)}, A\mathbf{v}^{(k-1)} \rangle}.$$

También se puede mostrar que con esta elección de s_{k-1} tenemos $\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(i)} \rangle = 0$. Para cada $i = 1, 2, \ldots, k-2$ [Lu, p. 245]. Así, $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}\}$ es un conjunto A-ortogonal.

Una vez elegido $\mathbf{v}^{(k)}$, calculamos

$$t_k = \frac{\langle \mathbf{v}^{(k)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{r}^{(k-1)} + s_{k-1}\mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$

$$= \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} + s_{k-1}\frac{\langle \mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}.$$

Por el teorema 7.33, $\langle \mathbf{v}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle = 0$, de modo que

$$t_k = \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}. \tag{7.29}$$

Así,

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)}.$$

Para calcular $\mathbf{r}^{(k)}$, multiplicamos por A y restamos $\mathbf{b}$ para obtener

$$A\mathbf{x}^{(k)} - \mathbf{b} = A\mathbf{x}^{(k-1)} - \mathbf{b} + t_k A\mathbf{v}^{(k)}$$

o bien

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - t_k A\mathbf{v}^{(k)}.$$

Así,

$$\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle = \langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k)} \rangle - t_k \langle A\mathbf{v}^{(k)}, \mathbf{r}^{(k)} \rangle = -t_k \langle \mathbf{r}^{(k)}, A\mathbf{v}^{(k)} \rangle.$$

Además, de la ecuación (7.29),

$$\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle = t_k \langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle,$$

de modo que

$$s_k = -\frac{\langle \mathbf{v}^{(k)}, A\mathbf{r}^{(k)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle} = -\frac{\langle \mathbf{r}^{(k)}, A\mathbf{v}^{(k)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}$$

$$= \frac{(1/t_k)\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{(1/t_k)\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle} = \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}.$$

En resumen, tenemos las fórmulas

$$\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}; \qquad \mathbf{v}^{(1)} = \mathbf{r}^{(0)};$$

y para $k = 1, 2, \ldots, n$,

$$\begin{aligned}
t_k &= \frac{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle},\\
\mathbf{x}^{(k)} &= \mathbf{x}^{(k-1)} + t_k \mathbf{v}^{(k)},\\
\mathbf{r}^{(k)} &= \mathbf{r}^{(k-1)} - t_k A\mathbf{v}^{(k)},\\
s_k &= \frac{\langle \mathbf{r}^{(k)}, \mathbf{r}^{(k)} \rangle}{\langle \mathbf{r}^{(k-1)}, \mathbf{r}^{(k-1)} \rangle},\\
\mathbf{v}^{(k+1)} &= \mathbf{r}^{(k)} + s_k \mathbf{v}^{(k)}.
\end{aligned} \tag{7.30}$$

En vez de presentar un algoritmo para el método del gradiente conjugado mediante estas fórmulas, ampliaremos el método para incluir el *precondicionamiento*. Si la matriz A es mal condicionada, este método es altamente susceptible a los errores de redondeo. Así, aunque la respuesta exacta debería obtenerse en n pasos, esto no es usual. Como método directo, el método del gradiente conjugado no es tan bueno como la eliminación gaussiana con pivoteo. El principal uso del método del gradiente conjugado es como método iterativo

aplicado a un sistema mejor condicionado. En este caso, con frecuencia se obtiene una solución aproximada aceptable en cerca de $\sqrt{n}$ pasos.

Para aplicar el método a un sistema mejor condicionado, queremos elegir una matriz no singular de condicionamiento C de modo que

$$\tilde{A} = C^{-1}A(C^{-1})^t$$

esté mejor condicionada. Para simplificar la notación, usaremos la matriz C^{-t} para referirnos a $(C^{-1})^t$.

Consideremos el sistema lineal

$$\tilde{A}\,\tilde{\mathbf{x}} = \tilde{\mathbf{b}},$$

donde $\tilde{\mathbf{x}} = C^t\,\mathbf{x}$ y $\tilde{\mathbf{b}} = C^{-1}\mathbf{b}$. Entonces

$$\tilde{A}\,\tilde{\mathbf{x}} = (C^{-1}AC^{-t})\,(C^t\mathbf{x}) = C^{-1}A\mathbf{x}.$$

Así, podríamos resolver $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ en términos de $\tilde{\mathbf{x}}$ y luego obtener $\mathbf{x}$ multiplicando por C^{-t}. Sin embargo, en vez de escribir de nuevo las ecuaciones (7.30) usado $\tilde{\mathbf{r}}^{(k)}$, $\tilde{\mathbf{v}}^{(k)}$, $\tilde{t}_k$, $\tilde{\mathbf{x}}^{(k)}$ y s_k, incorporamos el precondicionamiento de manera implícita.

Como

$$\tilde{\mathbf{x}}^{(k)} = C^t\mathbf{x}^{(k)},$$

tenemos

$$\tilde{\mathbf{r}}^{(k)} = \tilde{\mathbf{b}} - \tilde{A}\,\tilde{\mathbf{x}}^{(k)} = C^{-1}\,\mathbf{b} - (C^{-1}AC^{-t})C^t\mathbf{x}^{(k)} = C^{-1}\,(\mathbf{b} - A\mathbf{x}^{(k)}) = C^{-1}\,\mathbf{r}^{(k)}.$$

Sean $\tilde{\mathbf{v}}^{(k)} = C^t\,\mathbf{v}^{(k)}$ y $\mathbf{w}^{(k)} = C^{-1}\,\mathbf{r}^{(k)}$. Entonces

$$\tilde{s}_k = \frac{\langle \tilde{\mathbf{r}}^{(k)}, \tilde{\mathbf{r}}^{(k)} \rangle}{\langle \tilde{\mathbf{r}}^{(k-1)}, \tilde{\mathbf{r}}^{(k-1)} \rangle} = \frac{\langle C^{-1}\mathbf{r}^{(k)}, C^{-1}\mathbf{r}^{(k)} \rangle}{\langle C^{-1}\mathbf{r}^{(k-1)}, C^{-1}\mathbf{r}^{(k-1)} \rangle},$$

de modo que

$$\tilde{s}_k = \frac{\langle \mathbf{w}^{(k)}, \mathbf{w}^{(k)} \rangle}{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}. \tag{7.31}$$

Así,

$$\tilde{t}_k = \frac{\langle \tilde{\mathbf{r}}^{(k-1)}, \tilde{\mathbf{r}}^{(k-1)} \rangle}{\langle \tilde{\mathbf{v}}^{(k)}, \tilde{A}\tilde{\mathbf{v}}^{(k)} \rangle} = \frac{\langle C^{-1}\mathbf{r}^{(k-1)}, C^{-1}\mathbf{r}^{(k-1)} \rangle}{\langle C^t\,\mathbf{v}^{(k)}, C^{-1}AC^{-t}C^t\,\mathbf{v}^{(k)} \rangle} = \frac{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}{\langle C^t\,\mathbf{v}^{(k)}, C^{-1}A\mathbf{v}^{(k)} \rangle}$$

y

$$\tilde{t}_k = \frac{\langle \mathbf{w}^{(k-1)}, \mathbf{w}^{(k-1)} \rangle}{\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k)} \rangle}. \tag{7.32}$$

Además,

$$\tilde{\mathbf{x}}^{(k)} = \tilde{\mathbf{x}}^{(k-1)} + \tilde{t}_k\tilde{\mathbf{v}}^{(k)}, \quad \text{de modo que} \quad C^t\mathbf{x}^{(k)} = C^t\mathbf{x}^{(k-1)} + \tilde{t}_k\,C^t\mathbf{v}^{(k)}$$

y

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \tilde{t}_k\mathbf{v}^{(k)}. \tag{7.33}$$

Continuando de esta forma,

$$\tilde{\mathbf{r}}^{(k)} = \tilde{\mathbf{r}}^{(k-1)} - \tilde{t}_k \tilde{A} \tilde{\mathbf{v}}^{(k)},$$

de modo que

$$C^{-1}\mathbf{r}^{(k)} = C^{-1}\mathbf{r}^{(k-1)} - \tilde{t}_k C^{-1} A C^{-t} \tilde{v}^{(k)}, \qquad \mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \tilde{t}_k A C^{-t} C^t \mathbf{v}^{(k)},$$

y

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \tilde{t}_k A \mathbf{v}^{(k)}. \tag{7.34}$$

Por último,

$$\tilde{\mathbf{v}}^{(k+1)} = \tilde{\mathbf{r}}^{(k)} + \tilde{s}_k \tilde{\mathbf{v}}^{(k)} \qquad \text{y} \qquad C^t\mathbf{v}^{(k+1)} = C^{-1}\mathbf{r}^{(k)} + \tilde{s}_k C^t \mathbf{v}^{(k)},$$

y así

$$\mathbf{v}^{(k+1)} = C^{-t}C^{-1}\mathbf{r}^{(k)} + \tilde{s}_k \mathbf{v}^{(k)} = C^{-t}\mathbf{w}^{(k)} + \tilde{s}_k \mathbf{v}^{(k)}. \tag{7.35}$$

El método del gradiente conjugado precondicionado se basa en el uso de las ecuaciones (7.31) a (7.35) en el orden (7.32), (7.33), (7.34), (7.31), (7.35). El algoritmo 7.5 implanta este procedimiento.

ALGORITMO 7.5

Método del gradiente conjugado precondicionado

Para resolver $A\mathbf{x} = \mathbf{b}$ dada la matriz de precondicionamiento C^{-1} y la aproximación inicial $\mathbf{x}^{(0)}$:

ENTRADA el número de ecuaciones e incógnitas n; las entradas a_{ij}, $1 \le i, j \le n$ de la matriz A; las entradas b_j, $1 \le j \le n$ del vector $\mathbf{b}$; las entradas γ_{ij}, $1 \le i, j \le n$ de la matriz de precondicionamiento C^{-1}, las entradas x_i, $1 \le i \le n$ de la aproximación inicial $\mathbf{x} = \mathbf{x}^{(0)}$, el número máximo de iteraciones N; tolerancia TOL.

SALIDA la solución aproximada $x_1, \ldots x_n$ y el residual $r_1, \ldots r_n$ o un mensaje indicando que ha excedido el número de iteraciones

Paso 1 Tome $\mathbf{r} = \mathbf{b} - A\mathbf{x}$; (*Calcule* $\mathbf{r}^{(0)}$.)
$\mathbf{w} = C^{-1}\mathbf{r}$; (*Nota:* $\mathbf{w} = \mathbf{w}^{(0)}$)
$\mathbf{v} = C^{-t}\mathbf{w}$; (*Nota:* $\mathbf{v} = \mathbf{v}^{(1)}$)
$\alpha = \sum_{j=1}^{n} w_j^2$

Paso 2 Tome $k = 1$.

Paso 3 Mientras ($k \le N$) haga los pasos 4-7.

Paso 4 Si $\|\mathbf{v}\| < TOL$, entonces
SALIDA ('Vector solución'; $x_1, \ldots, x_n$);
SALIDA ('con residual'; $r_1, \ldots, x_n$);
(*El procedimiento tuvo éxito.*)
PARAR.

Paso 5 Tome $\mathbf{u} = A\mathbf{v}$; (*Nota:* $\mathbf{u} = A\mathbf{v}^{(k)}$)

$$t = \frac{\alpha}{\sum_{j=1}^{n} v_j u_j}; \ (\textit{Nota: } t = t_k)$$

$\mathbf{x} = \mathbf{x} + t\,\mathbf{v}$; (*Nota:* $\mathbf{x} = \mathbf{x}^{(k)}$)
$\mathbf{r} = \mathbf{r} - t\,\mathbf{u}$; (*Nota:* $\mathbf{r} = \mathbf{r}^{(k)}$)
$\mathbf{w} = C^{-1}\mathbf{r}$; (*Nota:* $\mathbf{w} = \mathbf{w}^{(k)}$)
$\beta = \sum_{j=1}^{n} w_j^2$. (*Nota:* $\beta = (\mathbf{w}^{(k)}, \mathbf{w}^{(k)})$)

Paso 6 Si $|\beta| < TOL$ entonces
si $\|\mathbf{r}\| < TOL$ entonces
SALIDA ('Vector solución'; $x_1, \ldots, x_n$);
SALIDA('con residual'; $r_1, \ldots, r_n$);
(*El procedimiento tuvo éxito*).
PARAR.

Paso 7 Tome $s = \beta/\alpha$; ($s = s_k$)
$\mathbf{v} = C^{-t}\mathbf{w} + s\mathbf{v}$; (*Nota:* $\mathbf{v} = \mathbf{v}^{(k+1)}$)
$\alpha = \beta$; (*Actualice* α.)
$k = k + 1$.

Paso 8 Si ($k > n$) entonces
SALIDA ('Se ha excedido el número máximo de iteraciones.');
(*El procedimiento no tuvo éxito.*)
PARAR. ■

El siguiente ejemplo ilustra los cálculos en un problema sencillo.

EJEMPLO 2 El sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned} 4x_1 + 3x_2 \phantom{{}- x_3} &= 24, \\ 3x_1 + 4x_2 - x_3 &= 30, \\ - x_2 + 4x_3 &= -24 \end{aligned}$$

tiene la solución $(3, 4, -5)^t$ y fue considerada en el ejemplo 3 de la sección 7.3. En ese ejemplo se usaron el método de Gauss-Seidel y el método SOR. Usaremos el método del gradiente conjugado sin precondicionamiento, de modo que $C = C^{-1} = I$. Sea $\mathbf{x}^{(0)} = (0, 0, 0)^t$. Entonces

$$\begin{aligned} \mathbf{r}^{(0)} &= \mathbf{b} - A\mathbf{x}^{(0)} = \mathbf{b} = (24, 30, -24)^t; \\ \mathbf{w} &= C^{-1}\mathbf{r}^{(0)} = (24, 30, -24)^t; \\ \mathbf{v}^{(1)} &= C^{-t}\mathbf{w} = (24, 30, -24)^t; \\ \alpha &= (\mathbf{w}, \mathbf{w}) = 2052. \end{aligned}$$

Comenzamos la primera iteración con $k = 1$. Entonces

$$\mathbf{u} = A\mathbf{v}^{(1)} = (186.0, 216.0, -126.0)^t;$$

$$t_1 = \frac{\alpha}{(\mathbf{v}^{(1)}, \mathbf{u})} = 0.1469072165;$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t_1\mathbf{v}^{(1)} = (3.525773196, 4.407216495, -3.525773196)^t;$$

$$\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - t_1\mathbf{u} = (-3.32474227, -1.73195876, -5.48969072)^t;$$
$$\mathbf{w} = C^{-1}\mathbf{r}^{(1)} = \mathbf{r}^{(1)};$$
$$\beta = \langle \mathbf{w}, \mathbf{w} \rangle = 44.19029651;$$
$$s_1 = \frac{\beta}{\alpha} = 0.02153523222;$$
$$\mathbf{v}^{(2)} = C^{-t}\mathbf{w} + s_1\mathbf{v}^{(1)} = (-2.807896697, -1.085901793, -6.006536293)^t;$$

Tome

$$\alpha = \beta = 44.19029651.$$

Ahora podemos iniciar la segunda iteración. Tenemos

$$\mathbf{u} = A\mathbf{v}^{(2)} = (-14.48929217, -6.760760967, -22.94024338)^t;$$
$$t_2 = 0.2378157558;$$
$$\mathbf{x}^{(2)} = (2.858011121, 4.148971939, -4.954222164)^t;$$
$$\mathbf{r}^{(2)} = (0.121039698, -0.124143281, -0.034139402)^t;$$
$$\mathbf{w} = C^{-1}\mathbf{r}^{(2)} = \mathbf{r}^{(2)};$$
$$\beta = 0.03122766148;$$
$$s_2 = 0.0007066633163;$$
$$\mathbf{v}^{(3)} = (0.1190554504, -0.1249106480, -0.03838400086)^t.$$

Tome

$$\alpha = \beta = 0.03122766148.$$

Por último, la tercera iteración da

$$\mathbf{u} = A\mathbf{v}^{(3)} = (0.1014898976, -0.1040922099, -0.0286253554)^t;$$
$$t_3 = 1.192628008;$$
$$\mathbf{x}^{(3)} = (2.999999998, 4.000000002, -4.999999998)^t;$$
$$\mathbf{r}^{(3)} = (0.36 \times 10^{-8}, 0.39 \times 10^{-8}, -0.141 \times 10^{-8})^t.$$

Como $\mathbf{x}^{(3)}$ es casi la solución exacta, el error de redondeo no afectó de manera significativa el resultado. En el ejemplo 3 de la sección 7.3, el método de Gauss-Seidel requirió 34 iteraciones, y el método SOR, con $\omega = 1.25$, requirió 14 iteraciones para una precisión de 10^{-7}. Sin embargo, debemos destacar que en este ejemplo estamos comparando en realidad un método directo con métodos iterativos. ■

El siguiente ejemplo ilustra el efecto del precondicionamento en una matriz pobremente condicionada. En este ejemplo y en lo sucesivo, $D^{-1/2}$ representará la matriz dia-

gonal cuyas entradas son los recíprocos de las raíces cuadradas de las entradas diagonales de la matriz de coeficientes A.

EJEMPLO 3 El sistema lineal $A\mathbf{x} = \mathbf{b}$ con

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

tiene la solución

$$\mathbf{x}^* = (7.859713071, 0.4229264082, -0.07359223906, -0.5406430164, 0.01062616286)^t.$$

La matriz A es simétrica y definida positiva, pero está mal condicionada, con número de condición $K_\infty(A) = 13961.71$. Usaremos la tolerancia 0.01 y compararemos los resultados obtenidos con los métodos iterativos de Jacobi, Gauss-Seidel y SOR (con $\omega = 1.25$) y con el método del gradiente conjugado, con $C^{-1} = I$. Luego establecemos el precondicionamiento eligiendo C^{-1} como $D^{-1/2}$, la matriz cuyas entradas diagonales son los recíprocos de las raíces cuadradas de las entradas diagonales de la matriz definida positiva A. Los resultados aparecen en la tabla 7.5. El método del gradiente conjugado precondicionado da la aproximación más precisa con el menor número de iteraciones. ■

Tabla 7.5

Método	Número de iteraciones	$\mathbf{x}^{(k)}$	$\|\mathbf{x}^* - \mathbf{x}^{(k)}\|_\infty$
Jacobi	49	$(7.86277141, 0.42320802, -0.07348669, -0.53975964, 0.01062847)^t$	0.00305834
Gauss-Seidel	15	$(7.83525748, 0.42257868, -0.07319124, -0.53753055, 0.01060903)^t$	0.02445559
SOR ($\omega = 1.25$)	7	$(7.85152706, 0.42277371, -0.07348303, -0.53978369, 0.01062286)^t$	0.00818607
Gradiente conjugado	5	$(7.85341523, 0.42298677, -0.07347963, -0.53987920, 0.008628916)^t$	0.00629785
Gradiente conjugado (Precondicionado)	4	$(7.85968827, 0.42288329, -0.07359878, -0.54063200, 0.01064344)^t$	0.00009312

Con frecuencia, el método del gradiente conjugado precondicionado se usa en la solución de sistemas lineales de gran tamaño, con una matriz esparcida y definida positiva. Estos sistemas deben resolverse para aproximar soluciones de problemas con valores en la frontera en ecuaciones diferenciales ordinarias (secciones 11.3, 11.4, 11.5). Mientras más grande sea el sistema, más impresionante será el método del gradiente conjugado, pues reduce de manera significativa el número de iteraciones necesarias. En estos sistemas, la matriz de precondicionamiento C es aproximadamente igual a L en la factorización de Choleski LL^t de A. Por lo general, las entradas pequeñas de A se ignoran y se aplica el método de Choleski para obtener lo que se llama una factorización incompleta LL^t de A. Así, $C^{-t}C^{-1} \approx A^{-1}$ y se obtiene una buena aproximación. En Kelley [Kelley] se puede obtener más información acerca del método del gradiente conjugado.

CONJUNTO DE EJERCICIOS 7.5

1. El sistema lineal

$$x_1 + \frac{1}{2}x_2 = \frac{5}{21},$$
$$\frac{1}{2}x_1 + \frac{1}{3}x_2 = \frac{11}{84}$$

tiene la solución $(x_1, x_2)^t = (1/6, 1/7)^t$.

a. Resuelva el sistema lineal mediante una eliminación gaussiana, con aritmética de redondeo a dos dígitos.

b. Resuelva el sistema lineal mediante el método del gradiente conjugado ($C = C^{-1} = I$) con una aritmética de redondeo a dos dígitos.

c. ¿Cuál método da la mejor respuesta?

d. Elija $C^{-1} = D^{-1/2}$. ¿Mejora esta elección al método del gradiente conjugado?

2. El sistema lineal

$$0.1x_1 + 0.2x_2 = 0.3,$$
$$0.2x_1 + 113x_2 = 113.2$$

tiene la solución $(x_1, x_2)^t = (1, 1)^t$. Repita las instrucciones del ejercicio 1 para este sistema lineal.

3. El sistema lineal

$$x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = \frac{5}{6},$$
$$\frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 = \frac{5}{12},$$
$$\frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = \frac{17}{60}$$

tiene la solución $(1, -1, 1)^t$.

a. Resuelva el sistema lineal mediante eliminación gaussiana, con aritmética de redondeo a tres dígitos.

b. Resuelva el sistema lineal mediante el método del gradiente conjugado con una aritmética de redondeo a tres dígitos.

c. ¿Mejora el pivoteo la respuesta en (a)?

d. Repita la parte (b) con $C^{-1} = D^{-1/2}$. ¿Mejora esto la respuesta en (b)?

4. Repita el ejercicio 3 usando la aritmética de precisión simple en una computadora.

5. Realice sólo dos pasos del método del gradiente conjugado con $C = C^{-1} = I$ en cada uno de los siguientes sistemas lineales. Compare los resultados de las partes (b), (c), (d) y (f) con los obtenidos en los ejercicios 1, 2 y 5 de la sección 7.3.

a. $$\begin{aligned} 3x_1 - x_2 + x_3 &= 1, \\ -x_1 + 6x_2 + 2x_3 &= 0, \\ x_1 + 2x_2 + 7x_3 &= 4, \end{aligned}$$

b. $$\begin{aligned} 10x_1 - x_2 &= 9, \\ -x_1 + 10x_2 - 2x_3 &= 7, \\ - 2x_2 + 10x_3 &= 6. \end{aligned}$$

c. $$\begin{aligned} 10x_1 + 5x_2 &= 6, \\ 5x_1 + 10x_2 - 4x_3 &= 25, \\ - 4x_2 + 8x_3 - x_4 &= -11, \\ - x_3 + 5x_4 &= -11. \end{aligned}$$

d. $$\begin{aligned} 4x_1 + x_2 - x_3 + x_4 &= -2, \\ x_1 + 4x_2 - x_3 - x_4 &= -1, \\ -x_1 - x_2 + 5x_3 + x_4 &= 0, \\ x_1 - x_2 + x_3 + 3x_4 &= 1. \end{aligned}$$

e. $$\begin{aligned} 4x_1 + x_2 + x_3 + \quad\quad x_5 &= 6, \\ x_1 + 3x_2 + x_3 + x_4 \quad\quad &= 6, \\ x_1 + x_2 + 5x_3 - x_4 - x_5 &= 6, \\ x_2 - x_3 + 4x_4 \quad\quad &= 6, \\ x_1 \quad - x_3 + \quad + 4x_5 &= 6. \end{aligned}$$

f. $$\begin{aligned} 4x_1 - x_2 \quad - x_4 \quad\quad &= 0, \\ -x_1 + 4x_2 - x_3 \quad - x_5 \quad &= 5, \\ - x_2 + 4x_3 \quad\quad - x_6 &= 0, \\ -x_1 \quad\quad + 4x_4 - x_5 \quad &= 6, \\ - x_2 \quad - x_4 + 4x_5 - x_6 &= -2, \\ - x_3 \quad - x_5 + 4x_6 &= 6. \end{aligned}$$

6. Repita el ejercicio 5 con $C^{-1} = D^{-1/2}$.
7. Repita el ejercicio 5 con $TOL = 10^{-3}$ en la norma l_∞. Compare los resultados de las partes (b), (c), (d) y (f) con los obtenidos en los ejercicios 3, 4 y 7 de la sección 7.3.
8. Repita el ejercicio 7 con $C^{-1} = D^{-1/2}$.
9. Use (i) el método de Jacobi, (ii) el método de Gauss-Seidel, (iii) el método SOR con $\omega = 1.3$ y (iv) el método del gradiente conjugado con precondicionamiento y $C^{-1} = D^{-1/2}$ para encontrar soluciones del sistema lineal $A\mathbf{x} = \mathbf{b}$ con una precisión de 10^{-5} en la norma l_∞.

a.

$$a_{i,j} = \begin{cases} 4, & \text{si} \quad j = i \text{ e } i = 1, 2, \ldots, 16, \\ -1, & \text{si} \begin{cases} j = i + 1 \text{ e } i = 1, 2, 3, 5, 6, 7, 9, 10, 11, 13, 14, 15, \\ j = i - 1 \text{ e } i = 2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16, \\ j = i + 4 \text{ e } i = 1, 2, \ldots, 12, \\ j = i - 4 \text{ e } i = 5, 6, \ldots, 16, \end{cases} \\ 0, & \text{en otro caso} \end{cases}$$

y

$$\mathbf{b} = (1.902207, 1.051143, 1.175689, 3.480083, 0.819600, -0.264419, -0.412789, 1.175689, 0.913337, -0.150209, -0.264419, 1.051143, 1.966694, 0.913337, 0.819600, 1.902207)^t$$

b.

$$a_{i,j} = \begin{cases} 4, & \text{si} \quad j = i \text{ e } i = 1, 2, \ldots, 25, \\ -1, & \text{si} \begin{cases} j = i + 1 \text{ e } i = \begin{cases} 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, \\ 16, 17, 18, 19, 21, 22, 23, 24, \end{cases} \\ j = i - 1 \text{ e } i = \begin{cases} 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, \\ 17, 18, 19, 20, 22, 23, 24, 25, \end{cases} \\ j = i + 5 \text{ e } i = 1, 2, \ldots, 20, \\ j = i - 5 \text{ e } i = 6, 7, \ldots, 25, \end{cases} \\ 0, & \text{en otro caso} \end{cases}$$

y

$$\mathbf{b} = (1, 0, -1, 0, 2, 1, 0, -1, 0, 2, 1, 0, -1, 0, 2, 1, 0, -1, 0, 2, 1, 0, -1, 0, 2)^t$$

c.

$$a_{i,j} = \begin{cases} 2i, & \text{si} \quad j = i \text{ e } i = 1, 2, \ldots, 40, \\ -1, & \text{si} \quad \begin{cases} j = i + 1 \text{ e } i = 1, 2, \ldots, 39, \\ j = i - 1 \text{ e } i = 2, 3, \ldots, 40, \end{cases} \\ 0, & \text{en otro caso} \end{cases}$$

y $b_i = 1.5i - 6$, para cada $i = 1, 2, \ldots, 40$.

10. Resuelva el sistema lineal del ejercicio 12(a) y (b) de la sección 7.3 usando el método del gradiente conjugado con $C^{-1} = I$.

11. Sean

$$A_1 = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix}, \quad -I = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix},$$

$$\text{y} \quad 0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Forme la matriz A de 16×16 con la siguiente partición.

$$A = \begin{bmatrix} A_1 & -I & 0 & 0 \\ -I & A_1 & -I & 0 \\ 0 & -I & A_1 & -I \\ 0 & 0 & -I & A_1 \end{bmatrix}.$$

Sea $\mathbf{b} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 1, 2, 3, 4, 5, 6)^t$.

a. Resuelva $A\mathbf{x} = \mathbf{b}$ usando el método del gradiente conjugado con tolerancia 0.05.

b. Resuelva $A\mathbf{x} = \mathbf{b}$ usando el método del gradiente conjugado precondicionado con $C^{-1} = D^{-1/2}$ y tolerancia 0.05.

c. ¿Existe alguna tolerancia para la cual los métodos de la parte (a) y la (b) requieran un número distinto de iteraciones?

12. Use las propiedades de la transpuesta dadas en el teorema 6.13 para demostrar el teorema 7.30.

13. a. Muestre que un conjunto A-ortogonal de vectores no nulos asociados con una matriz definida positiva es linealmente independiente.

b. Muestre que si $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ es un conjunto de vectores A-ortogonales no nulos en $\mathbb{R}$ y $\mathbf{z}^t\mathbf{v}^{(i)} = \mathbf{0}$, para cada $i = 1, 2, \ldots, n$, entonces $\mathbf{z} = \mathbf{0}$.

14. Demuestre el teorema 7.33 usando inducción matemática, como sigue:

a. Muestre que $\langle \mathbf{r}^{(1)}, \mathbf{v}^{(1)} \rangle = 0$.

b. Suponga que $\langle \mathbf{r}^{(k)}, \mathbf{v}^{(j)} \rangle = 0$, para cada $k \le l$ and $j = 1, 2, \ldots, k$ y muestre que esto implica que $\langle \mathbf{r}^{(l+1)}, \mathbf{v}^{(j)} \rangle = 0$, para cada $j = 1, 2, \ldots, l$.

c. Muestre que $\langle \mathbf{r}^{(l+1)}, \mathbf{v}^{(l+1)} \rangle = 0$.

7.6 Reseña de métodos y de software

En este capítulo estudiamos los métodos iterativos para aproximar la solución de los sistemas lineales. Comenzamos con el método de Jacobi y el de Gauss-Seidel, como una introducción a estos métodos. Ambas técnicas requieren una aproximación inicial arbitraria $\mathbf{x}^{(0)}$ y generan una secuencia de vectores $\mathbf{x}^{(i+1)}$ utilizando la ecuación de la forma

$$\mathbf{x}^{(i+1)} = T\mathbf{x}^{(i)} + \mathbf{c}.$$

Se señaló que el método convergerá si y sólo si el radio espectral de la matriz de iteración $\rho(T) < 1$ y que, cuanto más pequeño sea el radio espectral, más rápida será la convergencia. El análisis de los vectores residuales del método de Gauss-Seidel culminó en el método iterativo SOR, el cual involucra un parámetro ω que acelera la convergencia.

Estos métodos iterativos y modificaciones se emplean ampliamente en la resolución de los sistemas lineales que se presentan en la solución numérica de los problemas relacionados con el valor de frontera y de las ecuaciones diferenciales parciales (véanse los capítulos, 11 y 12). A menudo estos sistemas son extensos, con más de 10 000 ecuaciones con 10 000 incógnitas y se encuentran esparcidas con sus elementos distintos de cero en posiciones predecibles. Los métodos iterativos también son útiles en otros sistemas grandes esparcidos y es fácil adaptarlos para aprovecharlos eficientemente en las computadoras en paralelo.

Casi todos los paquetes comerciales y de dominio público que contienen métodos iterativos para resolver un sistema lineal de ecuaciones requieren un precondicionamiento para su método. Con frecuencia, se logra una convergencia más rápida de los casos iterativos introduciendo un precondicionamiento; éste produce un sistema equivalente de ecuaciones que, esperamos, exhiba mejores características de convergencia que el sistema original. La biblioteca IMSL tiene la subrutina PCGRC, que es un método de gradiente conjugado precondicionado. La biblioteca NAG tiene varias subrutinas con el prefijo F11 para resolver sistemas lineales en forma iterativa. Todas las subrutinas se basan en subespacios de Krylov. Saad [Sa2] tiene una descripción detallada de los métodos con subespacios de Krylov. Los paquetes LINPACK y LAPACK contienen sólo métodos directos para resolver sistemas lineales; sin embargo, los paquetes contienen muchas subrutinas que son utilizadas por los paquetes iterativos. Los paquetes de dominio público IML++, ITPACK, SLAP y Templates contienen métodos iterativos. MATLAB contiene varios métodos iterativos que también se basan en subespacios de Krylov. Por ejemplo, el comando $x = \text{PCG}(A, b)$ ejecuta el método del gradiente conjugado precondicionado para resolver el sistema lineal $Ax = b$. Algunos de los parámetros opcionales de entrada de PCG son: TOL, una tolerancia para la convergencia; MAXIT, el número máximo de iteraciones; y M para el precondicionamiento.

Los conceptos de número de condición y de matriz mal condicionadas se expusieron en la sección 7.4. Muchas de las subrutinas con que se resuelve un sistema lineal o se factoriza una matriz con la factorización LU contienen verificaciones para las matrices mal condicionadas y además ofrecen una estimación del número de condición.

La subrutina SGETRF de LAPACK factoriza la matriz real A en una factorización LU, proporcionando además el ordenamiento de renglones para la matriz de permutación P, donde $PA = LU$. La subrutina SGECON también ofrece el número de condición de A. LAPACK cuenta con otras subrutinas para matrices especiales; por ejemplo, SPOTRF efectúa la factorización de Choleski para una matriz definida, positiva A y SPOCON estima el recíproco del número de condición usando la factorización de Choleski calculada por SPOTRF.

La biblioteca IMSL tiene subrutinas que estiman el numero de condición. Por ejemplo, la subrutina LFCRG calcula la factorización LU; $PA = LU$ de la matriz A y además ofrece una estimación del número de condición. La biblioteca NAG dispone de subrutinas semejantes.

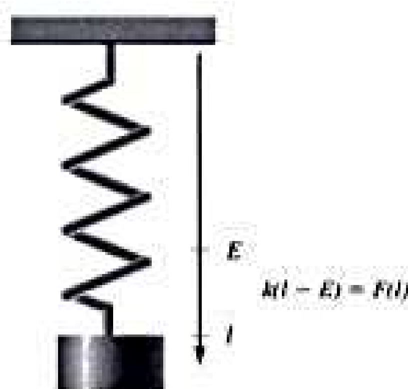
LAPACK, LINPACK, la biblioteca IMSL y la biblioteca NAG cuentan con subrutinas que mejoran una solución de un sistema lineal mal condicionado. Las subrutinas prueban el número de condición y después usan el refinamiento iterativo para obtener la solución más exacta posible con la precisión que ofrece la computadora.

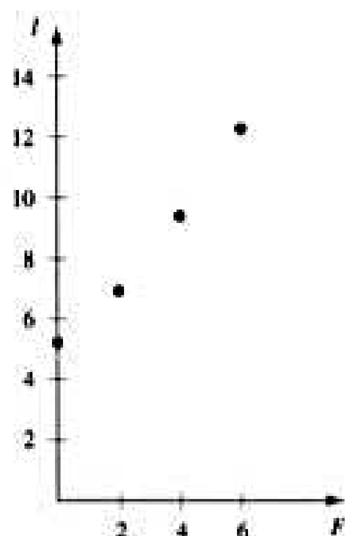
El lector que desee más información sobre los métodos iterativos con que se resuelven los sistemas lineales, puede consultar en Varga [Var], Young [Y], Hageman y Young [HY] y el libro reciente de Axelsson [Ax]. Los métodos iterativos para sistemas grandes esparcidos se describen en Barrett y otros autores [Barr], Hackbusch [Hac] Kelley [Kelley] y Saad [Sa2].

CAPÍTULO

8 Teoría de la aproximación

La ley de Hooke establece que, cuando se aplica una fuerza a un resorte fabricado de material uniforme, la longitud del resorte es una función lineal de la fuerza aplicada. Podemos escribir la función lineal como $F(l) = k(l - E)$, donde $F(l)$ representa la fuerza requerida para extender el resorte l unidades, la constante E representa la longitud del resorte sin que se aplique fuerza alguna y la constante k es la constante del resorte.





Supongamos que queremos determinar la constante de un resorte cuya longitud inicial es de 5.3 plg. Aplicamos al resorte las fuerzas de 2, 4 y 6 lb consecutivamente y observamos que su longitud aumenta a 7.0, 9.4 y 12.3 plg, respectivamente. Un análisis rápido muestra que los puntos (0, 5.3), (2, 7.0), (4, 9.4) y (6, 12.3) no se encuentran exactamente en una línea recta. Aunque podríamos utilizar un par aleatorio de estos puntos de datos para aproximar la



constante del resorte, parecería más razonable obtener la línea que aproxime *mejor* todos los puntos de datos para determinar la constante. En el capítulo que ahora nos ocupa estudiaremos este tipo de aproximación; la aplicación del resorte se incluye en el ejercicio 7 de la sección 8.1.

El estudio de la teoría de la aproximación comprende dos tipos generales de problemas. Uno se presenta cuando una función se da de manera explícita, pero queremos encontrar un tipo "más simple" de ella —un polinomio, por ejemplo— que nos sirva para determinar los valores aproximados de una función dada. El otro problema de la teoría se refiere a la adaptación de las funciones a ciertos datos y a la búsqueda de la función "óptima" en una clase que podamos emplear para representar los datos.

En el capítulo 3 nos ocupamos de ambos problemas. El polinomio de Taylor de grado n alrededor del número x_0 constituye una excelente aproximación a una función f diferenciable multiplicada por $(n + 1)$ en una pequeña vecindad de x_0. Los polinomios interpolantes de Lagrange o, más generalmente, los polinomios oscilantes se trataron como polinomios de aproximación y como polinomios para ajustar determinados datos. En ese capítulo explicamos los trazadores cúbicos, las limitaciones para estas técnicas, y también se tratan otras rutas de crecimiento.

8.1 Aproximación discreta por mínimos cuadrados

Consideremos el problema de estimar los valores de una función en puntos no tabulados, si contamos con los datos experimentales de la tabla 8.1.

Tabla 8.1

x_i	y_i	x_i	y_i
1	1.3	6	8.8
2	3.5	7	10.1
3	4.2	8	12.5
4	5.0	9	13.0
5	7.0	10	15.6

La figura 8.1 muestra una gráfica de los valores de la tabla 8.1. Esta gráfica sugiere que la relación real entre x y y es lineal. La razón probable de que ninguna recta se ajuste a estos datos es que éstos tienen cierto error: no es razonable exigir que la función de aproximación coincida exactamente con los datos. De hecho, tal función introduciría oscilaciones que no estaban presentes en un inicio. Por ejemplo, el polinomio de interpolación de noveno grado para los datos que aparecen en la figura 8.2 se obtiene en Maple mediante los comandos

```
>p:=interp([1,2,3,4,5,6,7,8,9,10],
        [1.3,3.5,4.2,5.0,7.0,8.8,10.1,12.5,13.0,15.6],x);
>plot({p},x=1..10);
```

Es claro que este polinomio es un mal predictor de la información entre varios de los puntos dato. Sería mejor encontrar la "mejor" (en cierto sentido) recta de aproximación, aunque ésta no coincida con los datos en punto alguno.

Figura 8.1

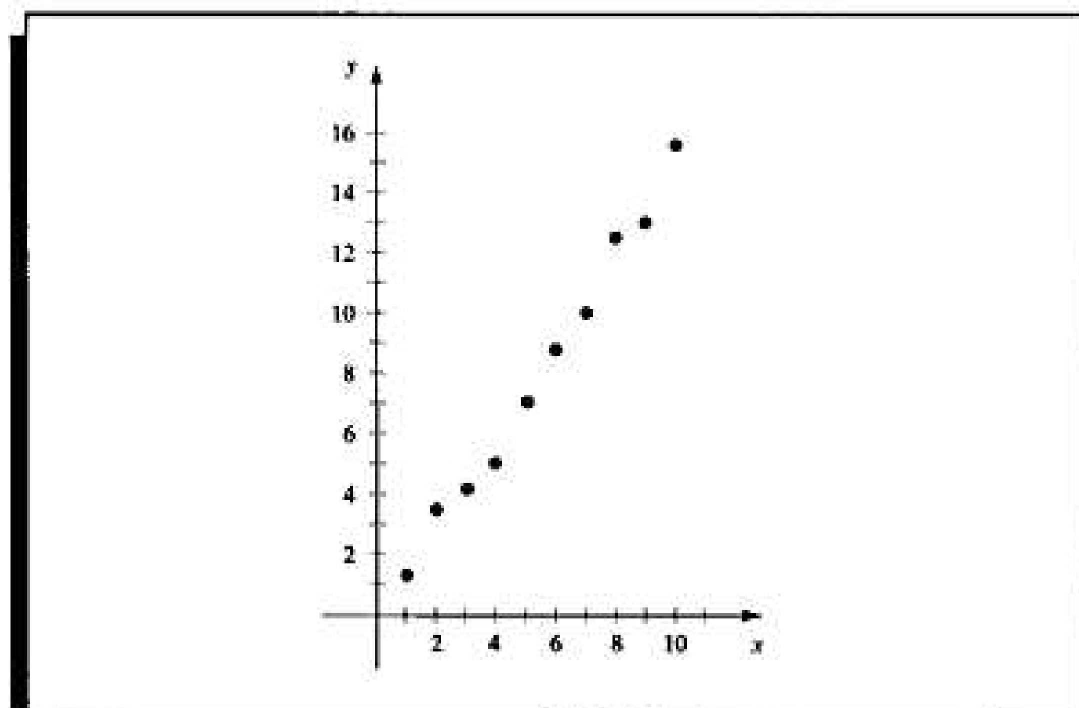
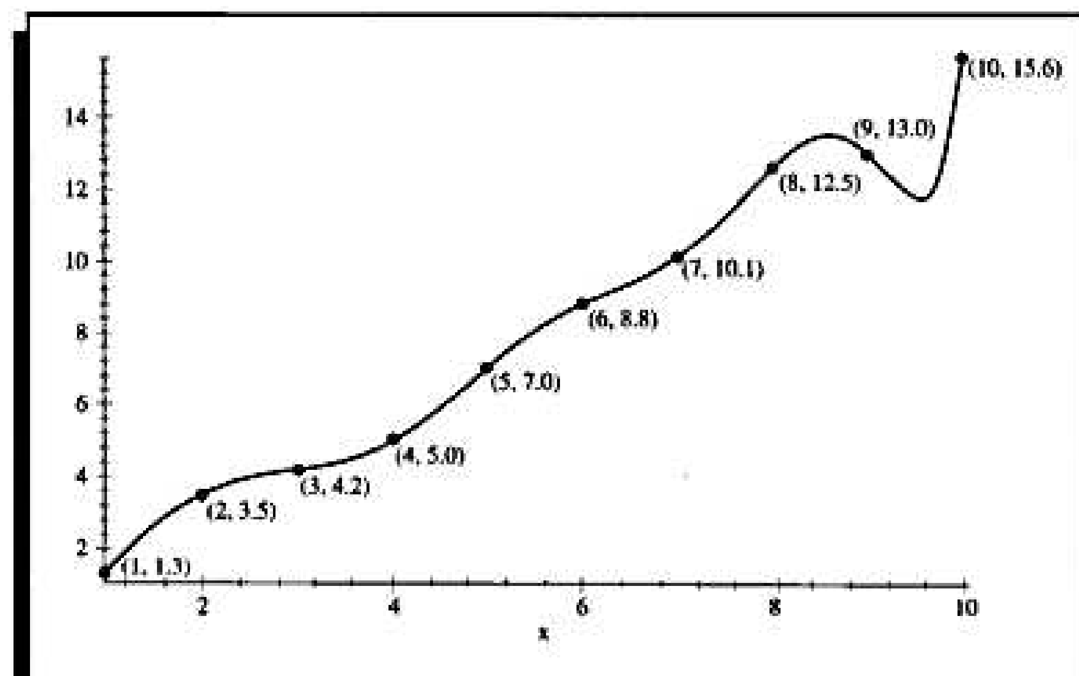


Figura 8.2



Sea $a_1x_i + a_0$ el i-ésimo valor de la recta de aproximación y y_i el i-ésimo valor dado para y. El problema de terminar la ecuación de la mejor aproximación lineal en el sentido absoluto consiste en hallar los valores de a_0 y a_1 que minimicen

$$E_\infty(a_0, a_1) = \max_{1 \leq i \leq 10} \{|y_i - (a_1x_i + a_0)|\}.$$

A lo anterior comúnmente se le llama problema **minimax**, y no se puede resolver mediante métodos elementales.

Otro método para determinar la mejor aproximación lineal implica hallar los valores de a_0 y a_1 que minimicen

$$E_1(a_0, a_1) = \sum_{i=1}^{10} |y_i - (a_1 x_i + a_0)|.$$

Esta cantidad se llama **desviación absoluta**. Para minimizar una función de dos variables, necesitamos igualar a cero sus derivadas parciales y resolver en forma simultánea las ecuaciones resultantes. En el caso de la desviación absoluta, necesitamos hallar a_0 y a_1 tales que

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{10} |y_i - (a_1 x_i + a_0)| \quad \text{y} \quad 0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{10} |y_i - (a_1 x_i + a_0)|.$$

La complejidad de este procedimiento radica en que la función valor absoluto no es derivable en cero, y no necesariamente se pueden obtener las soluciones de este par de ecuaciones.

El método de **mínimos cuadrados** para resolver este problema requiere determinar la mejor línea de aproximación, cuando el error es la suma de los cuadrados de las diferencias entre los valores de y en la línea aproximación y los valores de y dados. Por tanto, hay que encontrar las constantes a_0 y a_1 que reduzcan al mínimo el error de mínimos cuadrados:

$$E_2(a_0, a_1) = \sum_{i=1}^{10} [y_i - (a_1 x_i + a_0)]^2.$$

El método de mínimos cuadrados es el procedimiento más adecuado para determinar las mejores aproximaciones lineales, pero hay importantes consideraciones teóricas que lo favorecen. El método minimax generalmente le da demasiado valor relativo a un pequeño elemento de datos que contiene un gran error. El método que utiliza la desviación absoluta simplemente promedia el error en varios puntos, sin dar suficiente valor relativo a un punto que está muy alejado de la aproximación. El método de mínimos cuadrados concede mayor valor relativo al punto que está alejado del resto de los datos, pero no permitirá que ese punto domine enteramente la aproximación. Una razón más para explicar el método de mínimos cuadrados es estudiar la distribución estadística del error. (Véase [Lar, pp. 463-481].)

El problema general de ajustar la mejor recta con mínimos cuadrados a una colección de datos $\{(x_i, y_i)\}_{i=1}^{m}$ implica minimizar el error total,

$$E \equiv E_2(a_0, a_1) = \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2,$$

con respecto a los parámetros a_0 y a_1. Para que haya un mínimo, debemos tener

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} [(y_i - (a_1 x_i - a_0)]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-1)$$

y

$$0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2 = 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-x_i).$$

Estas ecuaciones se simplifican en las **ecuaciones normales**:

$$a_0 \cdot m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \quad \text{y} \quad a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i.$$

La solución de este sistema de ecuaciones es

$$a_0 = \frac{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2} \tag{8.1}$$

y

$$a_1 = \frac{m\sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2}. \tag{8.2}$$

EJEMPLO 1 Considere los datos de la tabla 8.1. Obtenga la línea de mínimos cuadrados que aproxima estos datos, extendiendo la tabla y sumando las columnas como se indica en las columnas tercera y cuarta de la tabla 8.2.

Tabla 8.2

x_i	y_i	x_i^2	$x_i y_i$	$P(x_i) = 1.538x_i - 0.360$
1	1.3	1	1.3	1.18
2	3.5	4	7.0	2.72
3	4.2	9	12.6	4.25
4	5.0	16	20.0	5.79
5	7.0	25	35.0	7.33
6	8.8	36	52.8	8.87
7	10.1	49	70.7	10.41
8	12.5	64	100.0	11.94
9	13.0	81	117.0	13.48
10	15.6	100	156.0	15.02
55	81.0	385	572.4	$E = \sum_{i=1}^{10}(y_i - P(x_i))^2 \approx 2.34$

Las ecuaciones normales (8.1) y (8.2) implican que

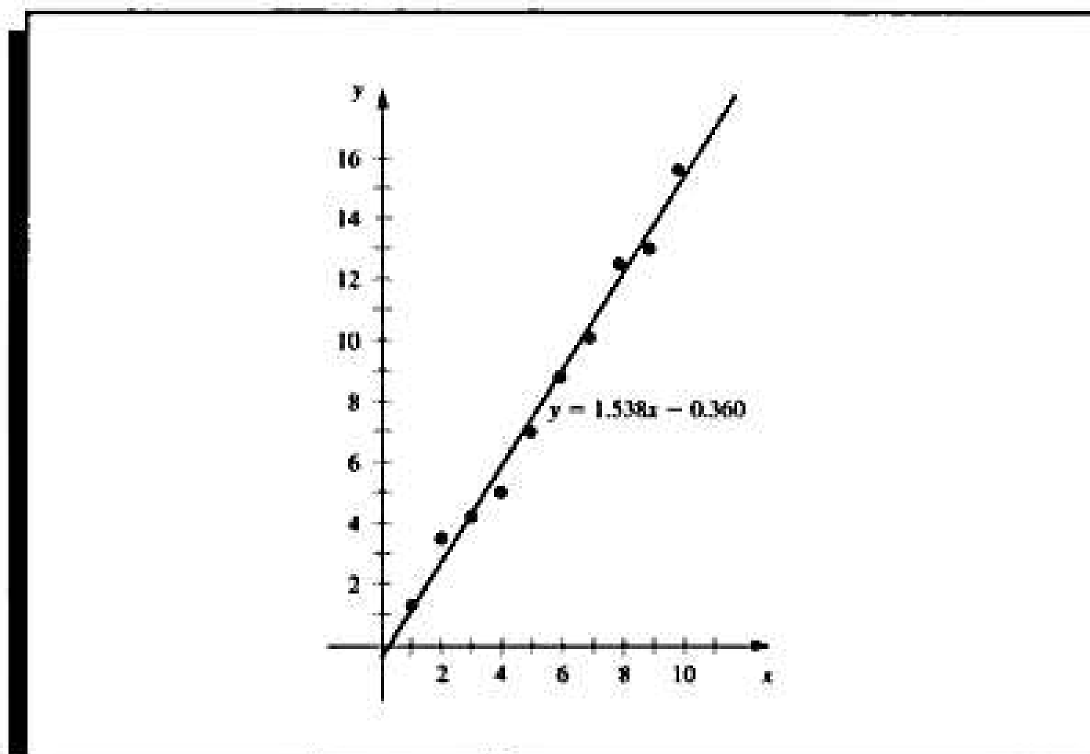
$$a_0 = \frac{385(81) - 55(572.4)}{10(385) - (55)^2} = -0.360$$

y que

$$a_1 = \frac{10(572.4) - 55(81)}{10(385) - (55)^2} = 1.538,$$

de modo que $P(x) = 1.538x - 0.360$. En la figura 8.3 se muestran la gráfica de esta línea y los puntos de datos. En la tabla 8.2 se incluyen los valores aproximados obtenidos con el método de mínimos cuadrados en los puntos de datos. ■

Figura 8.3



De modo similar se resuelve el problema general de aproximar un conjunto de datos $\{(x_i, y_i) \mid i = 1, 2, \ldots, m\}$, con un polinomio algebraico

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

De grado $n < m - 1$ mediante el procedimiento de mínimos cuadrados. Para disminuir al mínimo el error de mínimos cuadrados, es necesario seleccionar las constantes $a_0, a_1, \ldots, a_n$.

$$\begin{aligned}
E_2 &= \sum_{i=1}^{m}(y_i - P_n(x_i))^2 \\
&= \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m} P_n(x_i)y_i + \sum_{i=1}^{m}(P_n(x_i))^2 \\
&= \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right)y_i + \sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right)^2 \\
&= \sum_{i=1}^{m} y_i^2 - 2\sum_{j=0}^{n} a_j\left(\sum_{i=1}^{m} y_i x_i^j\right) + \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k\left(\sum_{i=1}^{m} x_i^{j+k}\right)
\end{aligned}$$

Igual que en el caso lineal, para reducir al mínimo E es necesario que $\partial E/\partial a_j = 0$ para cada $j = 0, 1, \ldots, n$. Así, para cada j,

$$0 = \frac{\partial E}{\partial a_j} = -2\sum_{i=1}^{m} y_i x_i^j + 2\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k}.$$

Esto nos da $n+1$ **ecuaciones normales** en las $n+1$ incógnitas a_j,

$$\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k} = \sum_{i=1}^{m} y_i x_i^j, \quad \text{para cada } j = 0, 1, \ldots, n. \tag{8.3}$$

Conviene escribir las ecuaciones como sigue:

$$a_0 \sum_{i=1}^{m} x_i^0 + a_1 \sum_{i=1}^{m} x_i^1 + a_2 \sum_{i=1}^{m} x_i^2 + \cdots + a_n \sum_{i=1}^{m} x_i^n = \sum_{i=1}^{m} y_i x_i^0,$$

$$a_0 \sum_{i=1}^{m} x_i^1 + a_1 \sum_{i=1}^{m} x_i^2 + a_2 \sum_{i=1}^{m} x_i^3 + \cdots + a_n \sum_{i=1}^{m} x_i^{n+1} = \sum_{i=1}^{m} y_i x_i^1,$$

$$\vdots$$

$$a_0 \sum_{i=1}^{m} x_i^n + a_1 \sum_{i=1}^{m} x_i^{n+1} + a_2 \sum_{i=1}^{m} x_i^{n+2} + \cdots + a_n \sum_{i=1}^{m} x_i^{2n} = \sum_{i=1}^{m} y_i x_i^n.$$

Puede demostrarse (véase el ejercicio 14) que las *ecuaciones normales* tienen una solución única, a condición de que las x_i sean distintas.

EJEMPLO 2 Ajuste los datos de la tabla 8.3 con el polinomio discreto de mínimos cuadrados de segundo grado. En este problema, $n = 2$, $m = 5$, y las tres ecuaciones normales son

$$\begin{aligned} 5a_0 + \quad 2.5a_1 + \quad 1.875a_2 &= 8.7680, \\ 2.5a_0 + \quad 1.875a_1 + 1.5625a_2 &= 5.4514, \\ 1.875a_0 + 1.5625a_1 + 1.3828a_2 &= 4.4015. \end{aligned}$$

Tabla 8.3

i	1	2	3	4	5
x_i	0.00	0.25	0.50	0.75	1.00
y_i	1.0000	1.2840	1.6487	2.1170	2.7183

Podemos resolver este sistema con un programa de álgebra por computadora. En Maple, primero definimos las ecuaciones

```
>eq1:=5*a0+2.5*a1+1.875*a2=8.7680;
>eq2:=2.5*a0+1.875*a1+1.5625*a2=5.4514;
>eq3:=1.875*a0+1.5625*a1+1.3828*a2=4.4015;
```

Para resolver el sistema hacemos

```
>solve({eq1,eq2,eq3},{a0,a1,a2});
```

lo que implica, con `Digits:=5;`

$$a_0 = 1.0051, \quad a_1 = 0.86468 \ \text{ y } \ a_2 = 0.84316.$$

Así pues, el polinomio de mínimos cuadrados de segundo grado que ajusta los datos anteriores es $P_2(x) = 1.0051 + 0.86468x + 0.84316x^2$, cuya gráfica aparece en la figura 8.4. En los valores dados de x_i tenemos las aproximaciones que se dan en la tabla 8.4.

El error total,

$$E_2 = \sum_{i=1}^{5} (y_i - P(x_i))^2 = 2.74 \times 10^{-4},$$

es el menor que puede obtenerse usando un polinomio que sea como máximo de segundo grado. ■

Figura 8.4

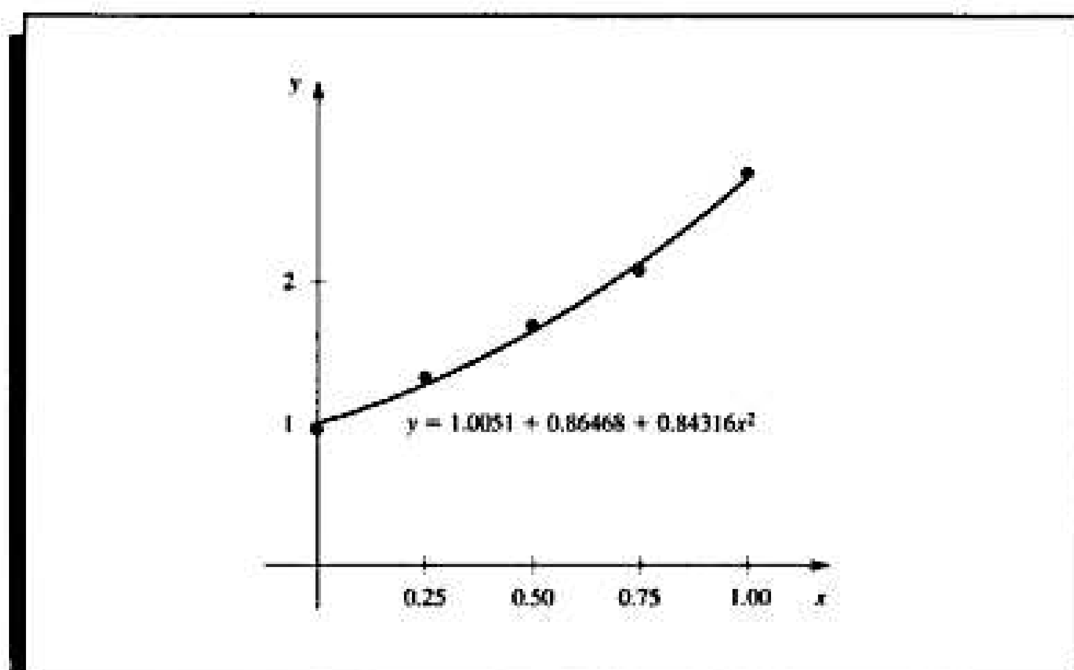


Tabla 8.4

i	1	2	3	4	5
x_i	0	0.25	0.50	0.75	1.00
y_i	1.0000	1.2840	1.6487	2.1170	2.7183
$P(x_i)$	1.0051	1.2740	1.6482	2.1279	2.7129
$y_i - P(x_i)$	−0.0051	0.0100	0.0004	−0.0109	0.0054

Maple tiene una función llamada `fit` en la biblioteca stats para calcular la aproximación del ejemplo 2 con el código de Maple:

```
>with(stats)
>xvals:=[0,0.25,0.5,0.75,1];
>yvals:=[1,1.284,1.6487,2.117,2.7183];
>z:=fit[leastsquare[[x,y],y=a*x^2+b*x+c,{a,b,c}]]
     ([xvals,yvals]);
```

Maple regresa el resultado

$$z := y = .8436571429x^2 + .8641828571x + 1.005137143$$

Para obtener una aproximación $y(1.7)$, escribimos

```
>evalf(subs(x=1.7,z));
```

para obtener $y = 4.912417143$.

Algunas veces conviene suponer que los datos tienen una relación exponencial. Para ello, la función de aproximación debe tener la forma

$$y = be^{ax} \tag{8.4}$$

o bien la forma

$$y = bx^a, \tag{8.5}$$

para algunas constantes a y b. El problema que implica aplicar el método de mínimos cuadrados en este tipo de situación se debe al intento de reducir al mínimo

$$E = \sum_{i=1}^{m}(y_i - be^{ax_i})^2, \quad \text{en el caso de la ecuación (8.4),}$$

o bien

$$E = \sum_{i=1}^{m}(y_i - bx_i^a)^2, \quad \text{en el caso de la ecuación (8.5).}$$

Las ecuaciones normales asociadas a estos procedimientos se obtienen de

$$0 = \frac{\partial E}{\partial b} = 2\sum_{i=1}^{m}(y_i - be^{ax_i})(-e^{ax_i})$$

y de

$$0 = \frac{\partial E}{\partial a} = 2\sum_{i=1}^{m}(y_i - be^{ax_i})(-bx_ie^{ax_i}), \text{ en el caso de la ecuación (8.4),}$$

o bien de

$$0 = \frac{\partial E}{\partial b} = 2\sum_{i=1}^{m}(y_i - bx_i^a)(-x_i^a)$$

y

$$0 = \frac{\partial E}{\partial a} = 2\sum_{i=1}^{m}(y_i - bx_i^a)(-b(\ln x_i)x_i^a) \text{ en el caso de la ecuación (8.5).}$$

Generalmente no es posible obtener una solución exacta para cualquiera de estos sistemas en a y en b.

El método que suele emplearse cuando se sospecha que los datos tienen una relación exponencial, consiste en considerar el logaritmo de la ecuación de aproximación:

$$\ln y = \ln b + ax, \quad \text{en el caso de la ecuación (8.4),}$$

y

$$\ln y = \ln b + a \ln x, \quad \text{en el caso de la ecuación (8.5).}$$

En uno y otro caso surge ahora un problema lineal, y las soluciones para $\ln b$ y a pueden obtenerse modificando debidamente las ecuaciones normales (8.1) y (8.2).

Sin embargo, la aproximación así obtenida *no* es la de mínimos cuadrados del problema original, y en algunos casos puede diferir significativamente de la aproximación de los mínimos cuadrados al problema original. En la aplicación del ejercicio 13 se describe este problema. Esa aplicación será reconsiderada como ejercicio 7 en la sección 10.3, donde la solución exacta del problema de los mínimos cuadrados se aproxima empleando los métodos adecuados para resolver los sistemas de ecuaciones no lineales.

EJEMPLO 3 Consideremos el conjunto de datos de las tres primeras columnas de la tabla 8.5.

Tabla 8.5

i	x_i	y_i	$\ln y_i$	x_i^2	$x_i \ln y_i$
1	1.00	5.10	1.629	1.0000	1.629
2	1.25	5.79	1.756	1.5625	2.195
3	1.50	6.53	1.876	2.2500	2.814
4	1.75	7.45	2.008	3.0625	3.514
5	2.00	8.46	2.135	4.0000	4.270
	7.50		9.404	11.875	14.422

Si graficamos x_i con $\ln y_i$, los datos parecen tener una relación lineal, de modo que es razonable suponer una aproximación de la forma

$$y = be^{ax} \quad \text{o} \quad \ln y = \ln b + ax.$$

Al extender la tabla y sumar las columnas apropiadas, obtenemos los datos residuales de la tabla 8.5.

Si utilizamos las ecuaciones normales (8.1) y (8.2),

$$a = \frac{(5)(14.422) - (7.5)(9.404)}{(5)(11.875) - (7.5)^2} = 0.5056$$

y

$$\ln b = \frac{(11.875)(9.404) - (14.422)(7.5)}{(5)(11.875) - (7.5)^2} = 1.122.$$

Puesto que $b = e^{1.122} = 3.071$, la aproximación adopta la forma

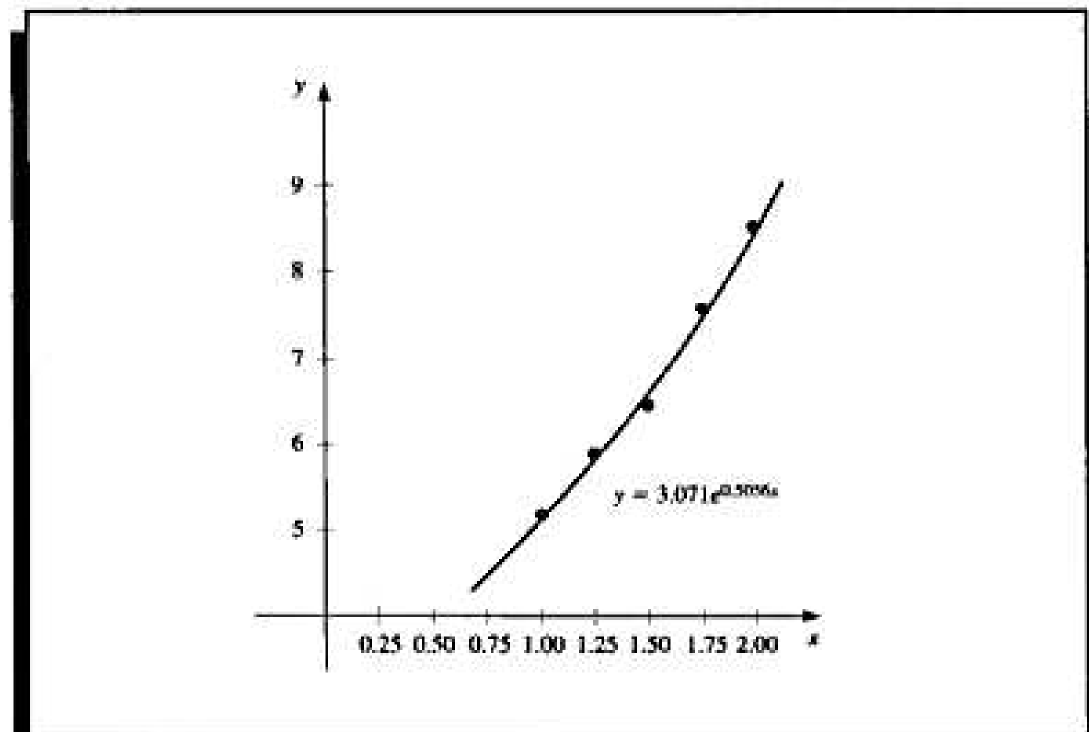
$$y = 3.071e^{0.5056x},$$

que, en los puntos de datos, proporciona los valores de la tabla 8.6 (véase la figura 8.5). ■

Tabla 8.6

i	x_i	y_i	$3.071e^{0.5056x_i}$
1	1.00	5.10	5.09
2	1.25	5.79	5.78
3	1.50	6.53	6.56
4	1.75	7.45	7.44
5	2.00	8.46	8.44

Figura 8.5



CONJUNTO DE EJERCICIOS 8.1

1. Calcule el polinomio lineal de mínimos cuadrados para los datos del ejemplo 2.
2. Calcule el polinomio de mínimos cuadrados de segundo grado para los datos del ejemplo 1 y compare el error total E en los dos polinomios.

3. Obtenga los polinomios de mínimos cuadrados de primero, segundo y tercer grados para los datos de la tabla anexa. En cada caso calcule el error E. Grafique los datos y los polinomios.

x_i	1.0	1.1	1.3	1.5	1.9	2.1
y_i	1.84	1.96	2.21	2.45	2.94	3.18

4. Obtenga los polinomios de mínimos cuadrados de primero, segundo y tercer grados para los datos de la tabla adjunta. En cada caso calcule el error E. Grafique los datos y los polinomios.

x_i	0	0.15	0.31	0.5	0.6	0.75
y_i	1.0	1.004	1.031	1.117	1.223	1.422

5. Con los datos

x_i	4.0	4.2	4.5	4.7	5.1	5.5	5.9	6.3	6.8	7.1
y_i	102.56	113.18	130.11	142.05	167.53	195.14	224.87	256.73	299.50	326.72

a. Construya el polinomio de mínimos cuadrados de primer grado y calcule el error.
b. Construya el polinomio de mínimos cuadrados de segundo grado y calcule el error.
c. Construya el polinomio de mínimos cuadrados de tercer grado y calcule el error.
d. Construya la aproximación de mínimos cuadrados de la forma be^{ax} y calcule el error.
e. Construya la aproximación de mínimos cuadrados de la forma bx^a y calcule el error.

6. Repita el ejercicio 5 con los siguientes datos.

x_i	0.2	0.3	0.6	0.9	1.1	1.3	1.4	1.6
y_i	0.050446	0.098426	0.33277	0.72660	1.0972	1.5697	1.8487	2.5015

7. En el primer ejemplo de este capítulo, describimos un experimento para determinar la constante de resorte k según la ley de Hooke:

$$F(l) = k(l - E).$$

La función F es la fuerza necesaria para alargar el resorte l unidades, donde la constante $E = 5.3$ plg, es la longitud del resorte sin estirar.

a. Suponga que se realizan mediciones de la longitud l, en pulgadas, para los pesos aplicados $F(l)$, en libras, como se indica en la tabla anexa:

$F(l)$	l
2	7.0
4	9.4
6	12.3

Obtenga la aproximación de mínimos cuadrados para k.

b. Se efectúan más mediciones, obteniéndose los datos adicionales:

$F(l)$	l
3	8.3
5	11.3
8	14.4
10	15.9

Calcule la nueva aproximación de mínimos cuadrados para k. Después, compare el error total en las partes (a) y (b).

8. La tabla siguiente contiene las calificaciones de las tareas escolares y las de los exámenes finales de 30 estudiantes de análisis numérico. Encuentre la ecuación de la recta de mínimos cuadrados para estos datos, y use esta recta para determinar la calificación de la tarea escolar que se requiere para predecir las calificaciones mínimas A (90%) y D (60%) en los exámenes finales.

Tarea	Examen final	Tarea	Examen final
302	45	323	83
325	72	337	99
285	54	337	70
339	54	304	62
334	79	319	66
322	65	234	51
331	99	337	53
279	63	351	100
316	65	339	67
347	99	343	83
343	83	314	42
290	74	344	79
326	76	185	59
233	57	340	75
254	45	316	45

9. La tabla siguiente muestra los promedios de puntos-calificación de 20 estudiantes universitarios de matemáticas y ciencias de la computación, junto con la puntuación que obtuvieron en la parte correspondiente a las matemáticas del examen American College Testing Program (programa de pruebas para las universidades estadounidenses), mientras cursaban la enseñanza media. Grafique los datos y obtenga para ellos una ecuación de recta de mínimos cuadrados.

Puntuación ACT	Promedio de grados de puntos	Puntuación ACT	Promedio de grados de puntos
28	3.84	29	3.75
25	3.21	28	3.65
28	3.23	27	3.87
27	3.63	29	3.75
28	3.75	21	1.66
33	3.20	28	3.12
28	3.41	28	2.96
29	3.38	26	2.92
23	3.53	30	3.10
27	2.03	24	2.81

10. El siguiente conjunto de datos, presentados al Senate Antitrust Subcommittee de Estados Unidos, muestra las características comparativas de choque-supervivencia de automóviles de varios tipos. Obtenga la recta de mínimos cuadrados que aproxima estos datos. (La tabla contiene el porcentaje de vehículos accidentados en los cuales la lesión más seria fue fatal o grave.)

Tipo	Peso promedio	Ocurrencia porcentual
1. Regular de lujo, de fabricación nacional	4800 lb	3.1
2. Regular intermedio, de fabricación nacional	3700 lb	4.0
3. Económico regular, de fabricación nacional	3400 lb	5.2
4. Compacto de fabricación nacional	2800 lb	6.4
5. Compacto de fabricación extranjera	1900 lb	9.6

11. Para determinar una relación entre el número de peces y el de las especies de peces en muestras tomadas de una parte de Great Barrier Reef, P. Sale y R. Dybdahl [SD]. Ajuste un polinomio lineal de mínimos cuadrados al siguiente conjunto de datos, que se recopilaron de muestras formadas durante un periodo de 2 años. Sea x el número de peces de la muestra y sea y el número de especies de la muestra.

x	y	x	y	x	y
13	11	29	12	60	14
15	10	30	14	62	21
16	11	31	16	64	21
21	12	36	17	70	24
22	12	40	13	72	17
23	13	42	14	100	23
25	13	55	22	130	34

Determine el polinomio lineal de mínimos cuadrados para estos datos.

12. Para determinar una relación funcional entre el coeficiente de atenuación y el grosor de una muestra de taconita, V. P. Singh [Si] ajusta un conjunto de datos por medio de un polinomio lineal de mínimos cuadrados. Los datos siguientes están tomados de una gráfica de ese trabajo. Obtenga el polinomio lineal de mínimos cuadrados que ajuste esos datos.

Espesor (cm)	Coeficiente de atenuación (dB/cm)
0.040	26.5
0.041	28.1
0.055	25.2
0.056	26.0
0.062	24.0
0.071	25.0
0.071	26.4
0.078	27.2
0.082	25.6
0.090	25.0
0.092	26.8

Espesor (cm)	Coeficiente de atenuación (dB/cm)
0.100	24.8
0.105	27.0
0.120	25.0
0.123	27.3
0.130	26.9
0.140	26.2

13. En un trabajo relacionado con el estudio de la eficiencia de la utilización de la energía por las larvas de la polilla modesta (*Pachysphinx modesta*), L. Schroeder [Schr1] utilizó los siguientes datos para determinar una relación entre W, el peso de las larvas vivas en gramos, y R, el consumo de oxígeno de las larvas en mililitros/hora. Por razones biológicas, se supone que entre W y R existe una relación de la forma $R = bW^a$.

a. Encuentre el polinomio logarítmico lineal de mínimos cuadrados usando

$$\ln R = \ln b + a \ln W.$$

b. Calcule el error asociado a la aproximación en la parte (a):

$$E = \sum_{i=1}^{37} (R_i - bW_i^a)^2.$$

c. Modifique la ecuación logarítmica de mínimos cuadrados de la parte (a) agregando el término cuadrático $c(\ln W_i)^2$, y después determine el polinomio logarítmico cuadrático de mínimos cuadrados.

d. Determine la fórmula y calcule el error asociado a la aproximación de la parte (c).

W	R	W	R	W	R	W	R	W	R
0.017	0.154	0.025	0.23	0.020	0.181	0.020	0.180	0.025	0.234
0.087	0.296	0.111	0.357	0.085	0.260	0.119	0.299	0.233	0.537
0.174	0.363	0.211	0.366	0.171	0.334	0.210	0.428	0.783	1.47
1.11	0.531	0.999	0.771	1.29	0.87	1.32	1.15	1.35	2.48
1.74	2.23	3.02	2.01	3.04	3.59	3.34	2.83	1.69	1.44
4.09	3.58	4.28	3.28	4.29	3.40	5.48	4.15	2.75	1.84
5.45	3.52	4.58	2.96	5.30	3.88			4.83	4.66
5.96	2.40	4.68	5.10					5.53	6.94

14. Demuestre que las ecuaciones normales (8.3) resultantes de la aproximación discreta de mínimos cuadrados producen una matriz simétrica no singular y que, por tanto, tienen una solución única. [*Sugerencia*: sea $A = (a_{ij})$, donde

$$a_{ij} = \sum_{k=1}^{m} x_k^{i+j-2}$$

y $x_1, x_2, \ldots, x_m$ son distintos con $n < m - 1$. Supongamos que A es singular y que $\mathbf{c} \neq \mathbf{0}$ es tal que $\mathbf{c}^t A \mathbf{c} = 0$. Demuestre que el polinomio de n-ésimo grado cuyos coeficientes son las coordenadas de $\mathbf{c}$ tiene más de n raíces; después use esto para establecer una contradicción.]

8.2 Polinomios ortogonales y aproximación por mínimos cuadrados

En la sección anterior estudiamos el problema de la aproximación por mínimos cuadrados para ajustar una colección de datos. El otro problema de aproximación mencionado en la introducción se refiere a la aproximación de funciones.

Supongamos que $f \in C[a, b]$ y que se requiere un polinomio $P_n(x)$ de grado máximo n, el cual reducirá al mínimo el error

$$\int_a^b [f(x) - P_n(x)]^2\, dx.$$

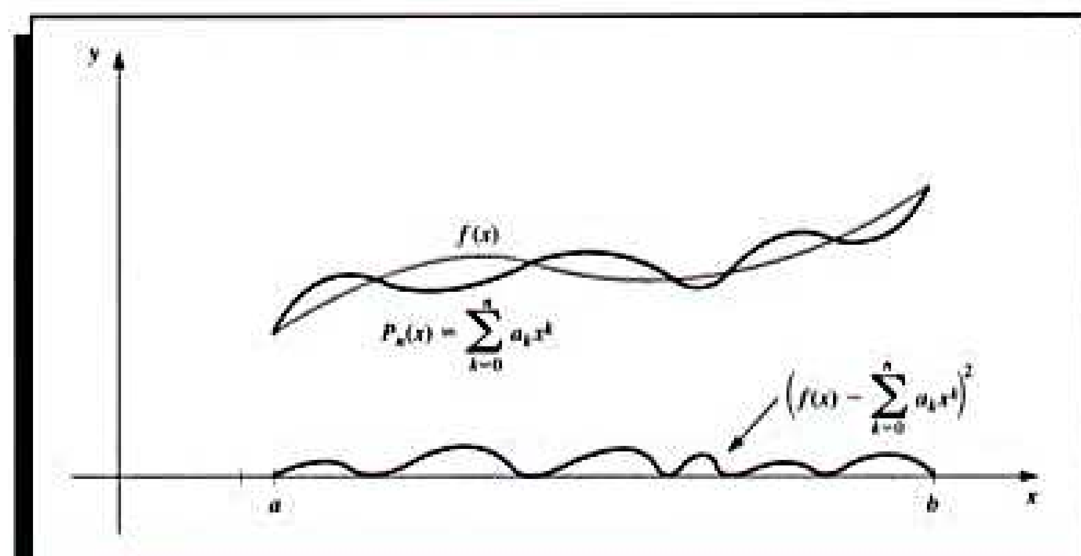
Si queremos determinar un polinomio de aproximación de mínimos cuadrados, esto es, un polinomio que reduzca al mínimo esta expresión, usamos

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = \sum_{k=0}^{n} a_k x^k,$$

y, como se observa en la figura 8.6, definimos

$$E = E(a_0, a_1, \ldots, a_n) = \int_a^b \left(f(x) - \sum_{k=0}^{n} a_k x^k\right)^2 dx.$$

Figura 8.6



El problema consiste en encontrar coeficientes reales $a_0, a_1, \ldots, a_n$ que reduzcan E al mínimo. Una condición necesaria para que los números $a_0, a_1, \ldots, a_n$ hagan eso es que

$$\frac{\partial E}{\partial a_j} = 0, \quad \text{para cada } j = 0, 1, \ldots, n.$$

Dado que

$$E = \int_a^b [f(x)]^2\,dx - 2\sum_{k=0}^{n} a_k \int_a^b x^k f(x)\,dx + \int_a^b \left(\sum_{k=0}^{n} a_k x^k\right)^2 dx,$$

tenemos

$$\frac{\partial E}{\partial a_j} = -2\int_a^b x^j f(x)\,dx + 2\sum_{k=0}^{n} a_k \int_a^b x^{j+k}\,dx.$$

Por tanto, si queremos obtener $P_n(x)$ debemos resolver las $(n + 1)$ **ecuaciones normales**

$$\sum_{k=0}^{n} a_k \int_a^b x^{j+k}\,dx = \int_a^b x^j f(x)\,dx, \quad \text{para cada } j = 0, 1, \ldots, n, \tag{8.6}$$

para las $(n + 1)$ incógnitas a_j. Puede demostrarse que las ecuaciones normales siempre tienen una solución única, siempre que $f \in C[a, b]$. (Véase el Ejercicio 15.)

EJEMPLO 1 Encuentre el polinomio de aproximación de mínimos cuadrados de segundo grado para la función $f(x) = \operatorname{sen} \pi x$ en el intervalo $[0, 1]$. Las ecuaciones normales para $P_2(x) = a_2x^2 + a_1x + a_0$ son

$$a_0\int_0^1 1\,dx + a_1\int_0^1 x\,dx + a_2\int_0^1 x^2\,dx = \int_0^1 \operatorname{sen}\pi x\,dx,$$

$$a_0\int_0^1 x\,dx + a_1\int_0^1 x^2\,dx + a_2\int_0^1 x^3\,dx = \int_0^1 x\operatorname{sen}\pi x\,dx,$$

$$a_0\int_0^1 x^2\,dx + a_1\int_0^1 x^3\,dx + a_2\int_0^1 x^4\,dx = \int_0^1 x^2\operatorname{sen}\pi x\,dx.$$

Al realizar la integración se obtiene

$$a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 = \frac{2}{\pi}, \quad \frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 = \frac{1}{\pi}, \quad \frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 = \frac{\pi^2 - 4}{\pi^3}.$$

Estas tres ecuaciones con tres incógnitas pueden resolverse para obtener

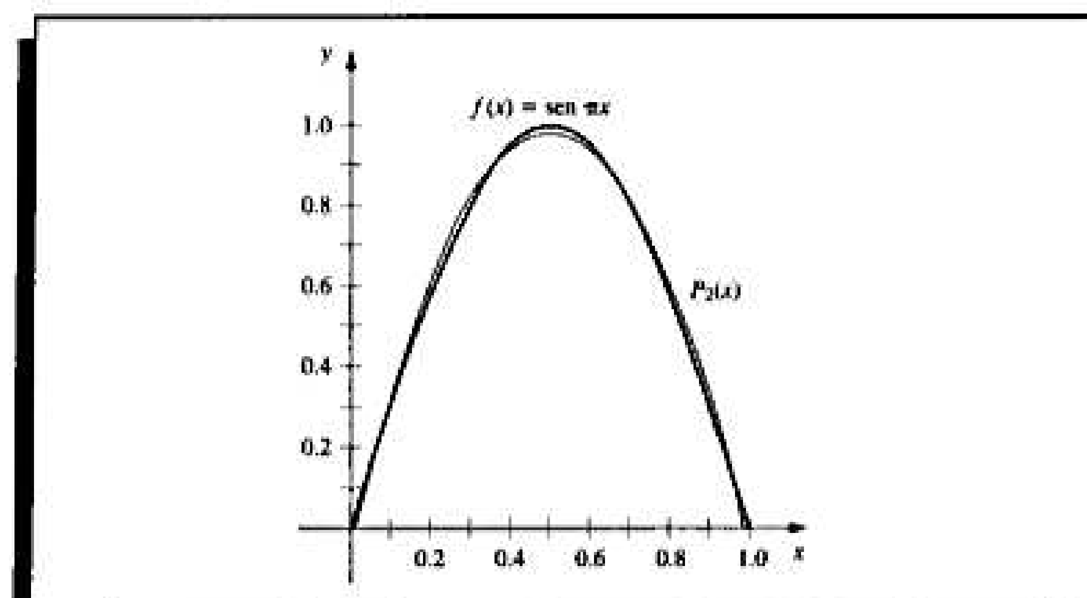
$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx -0.050465 \quad \text{y} \quad a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx 4.12251.$$

En consecuencia, la aproximación del polinomio de mínimos cuadrados de segundo grado para $f(x) = \operatorname{sen} \pi x$ en $[0, 1]$ es $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$. (Véase la figura 8.7.) ■

En el ejemplo 1 se ilustra el problema que implica obtener una aproximación del polinomio de mínimos cuadrados. Un sistema lineal $(n + 1) \times (n + 1)$ debe resolverse para las incógnitas $a_0, \ldots, a_n$. Los coeficientes del sistema lineal tienen la forma

$$\int_a^b x^{j+k}\,dx = \frac{b^{j+k+1} - a^{j+k+1}}{j + k + 1},$$

Figura 8.7



sistema lineal que no tiene una solución numérica que se pueda calcular fácilmente. La matriz del sistema lineal se conoce como **matriz de Hilbert**. Esta matriz mal condicionada constituye un ejemplo clásico para demostrar los problemas del error de redondeo. (Véase el ejercicio 6 de la sección 7.4.)

Otra desventaja se parece a la situación que se presentó cuando por primera vez tratamos los polinomios de Lagrange en la sección 3.1. Los cálculos efectuados para obtener el mejor polinomio de n-ésimo grado, $P_n(x)$, no reduce el trabajo que se requiere para calcular $P_{n+1}(x)$, o sea el polinomio del grado mayor siguiente.

A continuación explicaremos otra técnica con que se obtienen las aproximaciones de mínimos cuadrados. La técnica resulta ser eficiente desde el punto de vista de los cálculos a realizar; una vez que se conoce $P_n(x)$ es fácil determinar $P_{n+1}(x)$. Para facilitar la exposición necesitaremos aprender otros conceptos nuevos.

Definición 8.1 Se dice que el conjunto de funciones $\{\phi_0, \ldots, \phi_n\}$ es **linealmente independiente** en $[a, b]$ si, siempre que

$$c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0, \quad \text{para cualquier } x \in [a, b],$$

se tiene que $c_0 = c_1 = \cdots = c_n = 0$. De lo contrario se dice que el conjunto de funciones es **linealmente dependiente**. ■

Teorema 8.2 Si $\phi_j(x)$ es un polinomio de j-ésimo grado para cada $j = 0, 1, \ldots, n$, entonces $\{\phi_0, \ldots, \phi_n\}$ es linealmente independiente en cualquier intervalo $[a, b]$. ■

Demostración Supongamos que $c_0, \ldots, c_n$ son números reales para los cuales

$$P(x) = c_0\phi_0(x) + c_1\phi_1(x) + \cdots + c_n\phi_n(x) = 0, \quad \text{para cualquier } x \in [a, b].$$

Como el polinomio $P(x)$ se anula en $[a, b]$, los coeficientes de todas las potencias de x son cero. En particular, el coeficiente de x^n es cero. Y puesto que $c_n\phi_n(x)$ es el único término

de $P(x)$ que contiene x^n, hemos de tener $c_n = 0$ y

$$P(x) = \sum_{j=0}^{n-1} c_j \phi_j(x).$$

En esta representación de $P(x)$ el único término que contiene una potencia de x^{n-1} es $c_{n-1}\phi_{n-1}(x)$, de modo que este término también debe ser cero y

$$P(x) = \sum_{j=0}^{n-2} c_j \phi_j(x).$$

En forma similar, las constantes restantes $c_{n-2}, c_{n-3}, \ldots, c_1, c_0$ son todas cero, lo cual significa que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es linealmente independiente. ■ ■ ■

EJEMPLO 2 Sea $\phi_0(x) = 2$, $\phi_1(x) = x - 3$ y $\phi_2(x) = x^2 + 2x + 7$. De acuerdo con el teorema 8.2, $\{\phi_0, \phi_1, \phi_2\}$ es linealmente independiente en cualquier intervalo $[a, b]$. Supongamos que $Q(x) = a_0 + a_1x + a_2x^2$. Demostraremos que existen constantes c_0, c_1 y c_2 tales que $Q(x) = c_0\phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x)$. Nótese que

$$1 = \frac{1}{2}\phi_0(x), \quad x = \phi_1(x) + 3 = \phi_1(x) + \frac{3}{2}\phi_0(x),$$

y que

$$\begin{aligned} x^2 &= \phi_2(x) - 2x - 7 = \phi_2(x) - 2\left[\phi_1(x) + \frac{3}{2}\phi_0(x)\right] - 7\left[\frac{1}{2}\phi_0(x)\right] \\ &= \phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x). \end{aligned}$$

Por tanto,

$$\begin{aligned} Q(x) &= a_0\left[\frac{1}{2}\phi_0(x)\right] + a_1\left[\phi_1(x) + \frac{3}{2}\phi_0(x)\right] + a_2\left[\phi_2(x) - 2\phi_1(x) - \frac{13}{2}\phi_0(x)\right] \\ &= \left(\frac{1}{2}a_0 + \frac{3}{2}a_1 - \frac{13}{2}a_2\right)\phi_0(x) + [a_1 - 2a_2]\,\phi_1(x) + a_2\phi_2(x), \end{aligned}$$

de modo que cualquier polinomio cuadrático se puede expresar como una combinación lineal de $\phi_0(x)$, $\phi_1(x)$ y $\phi_2(x)$. ■

La situación descrita en el ejemplo 2 se cumple en un caso mucho más general. Sea Π_n **el conjunto de todos los polinomios de grado máximo n**. El siguiente resultado se utiliza ampliamente en muchas aplicaciones del álgebra lineal. En el ejercicio 13 veremos su comprobación.

Teorema 8.3 Si $\{\phi_0(x), \phi_1(x), \ldots, \phi_n(x)\}$ es un conjunto de polinomios linealmente independientes en Π_n, entonces todo polinomio en Π_n puede escribirse de manera única como una combinación lineal de $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$. ■

Para explicar la aproximación general de la función, es necesario presentar antes los conceptos de funciones de peso y ortogonalidad.

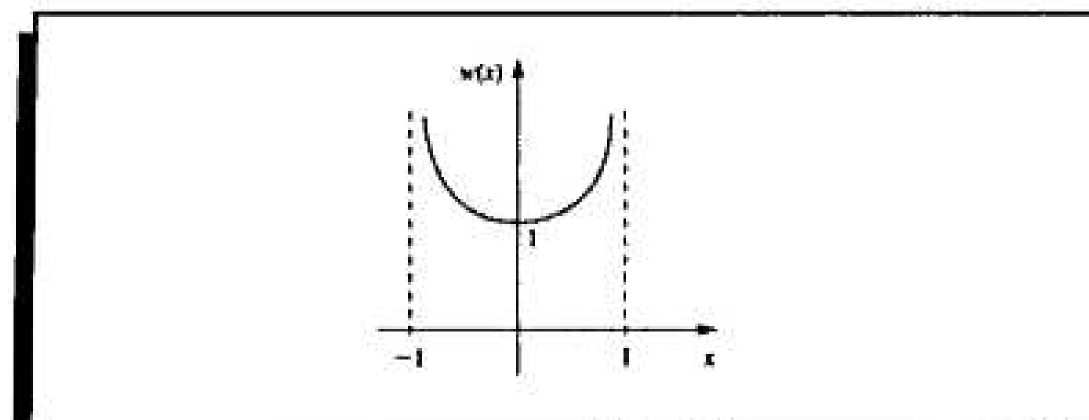
Definición 8.4 A una función integrable w se le llama **función de peso** en el intervalo I si $w(x) \geq 0$, para toda x en I, pero $w(x) \not\equiv 0$ en cualquier subintervalo de I. ■

Una función de peso tiene por objeto asignar diferentes grados de importancia a las aproximaciones en ciertas partes del intervalo. Por ejemplo, la función peso

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

pone menos énfasis cerca del centro del intervalo $(-1, 1)$ y mayor cuando $|x|$ se halla cerca del uno (véase la figura 8.8). Esta función de peso se empleará en la siguiente sección.

Figura 8.8



Supongamos que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un conjunto de funciones linealmente independientes en $[a, b]$, que w es una función de peso en $[a, b]$ y que, para $f \in C[a, b]$, se busca una combinación lineal

$$P(x) = \sum_{k=0}^{n} a_k \phi_k(x)$$

para reducir al mínimo el error

$$E(a_0, \ldots, a_n) = \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} a_k \phi_k(x) \right]^2 dx.$$

Este problema se reduce a la situación que se planteó al inicio de esta sección en el caso especial cuando $w(x) \equiv 1$ y $\phi_k(x) = x^k$, para cada $k = 0, 1, \ldots, n$.

Las ecuaciones normales asociadas a este problema se derivan del hecho de que para cada $j = 0, 1, \ldots, n$,

$$0 = \frac{\partial E}{\partial a_j} = 2 \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} a_k \phi_k(x) \right] \phi_j(x)\, dx.$$

El sistema de ecuaciones normales puede escribirse

$$\int_a^b w(x) f(x)\, \phi_j(x)\, dx = \sum_{k=0}^{n} a_k \int_a^b w(x)\, \phi_k(x)\phi_j(x)\, dx, \quad \text{para } j = 0, 1, \ldots, n.$$

Si podemos elegir las funciones $\phi_0, \phi_1, \ldots, \phi_n$ de modo que

$$\int_a^b w(x)\, \phi_k(x)\phi_j(x)\, dx = \begin{cases} 0, & \text{cuando } j \neq k, \\ \alpha_j > 0, & \text{cuando } j = k, \end{cases} \tag{8.7}$$

entonces las ecuaciones normales se reducen a

$$\int_a^b w(x) f(x)\phi_j(x)\, dx = a_j \int_a^b w(x)\, [\phi_j(x)]^2\, dx = a_j\alpha_j,$$

para cada $j = 0, 1, \ldots, n$ y fácilmente obtenemos

$$a_j = \frac{1}{\alpha_j} \int_a^b w(x) f(x)\phi_j(x)\, dx.$$

Por tanto, el problema de aproximación por mínimos cuadrados se simplifica considerablemente, cuando se seleccionan las funciones $\phi_0, \phi_1, \ldots, \phi_n$ que satisfagan la condición de *ortogonalidad* en la ecuación (8.7). En lo que resta de esta sección estudiaremos los conjuntos de este tipo.

Definición 8.5 Se dice que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un **conjunto ortogonal de funciones** en el intervalo $[a, b]$ respecto a la función de peso w, si

$$\int_a^b w(x)\phi_j(x)\phi_k(x)\, dx = \begin{cases} 0, & \text{cuando } j \neq k, \\ \alpha_k > 0, & \text{cuando } j = k. \end{cases}$$

Si además, $\alpha_k = 1$ para cada $k = 0, 1, \ldots, n$, se dice que el conjunto es **ortonormal**. ■

Esta definición, junto con los comentarios que la anteceden, da origen al siguiente teorema.

Teorema 8.6 Si $\{\phi_0, \ldots, \phi_n\}$ es un conjunto ortogonal de funciones en un intervalo $[a, b]$ respecto a la función de peso w, entonces la aproximación por mínimos cuadrados a f en $[a, b]$ respecto a w es

$$P(x) = \sum_{k=0}^{n} a_k\phi_k(x),$$

donde, para cada $k = 0, 1, \ldots, n$,

$$a_k = \frac{\int_a^b w(x)\phi_k(x) f(x)\, dx}{\int_a^b w(x)[\phi_k(x)]^2\, dx} = \frac{1}{\alpha_k} \int_a^b w(x)\phi_k(x) f(x)\, dx.$$ ■

Aunque la definición 8.5 y el teorema 8.6 son válidos para una clase amplia de funciones ortogonales, sólo tomaremos en cuenta conjuntos ortogonales de polinomios. El siguiente teorema, que tiene como fundamento el **proceso de Gram-Schmidt**, describe cómo construir polinomios ortogonales en $[a, b]$ respecto a la función de peso w.

Teorema 8.7 El conjunto de las funciones polinomiales $\{\phi_0, \phi_1, \ldots, \phi_n\}$ definida en la siguiente forma es ortogonal en $[a, b]$ respecto a la función de peso w.

$$\phi_0(x) = 1, \quad \phi_1(x) = x - B_1, \quad \text{para cada } x \text{ en } [a, b],$$

donde

$$B_1 = \frac{\int_a^b x w(x)[\phi_0(x)]^2\, dx}{\int_a^b w(x)[\phi_0(x)]^2\, dx},$$

y cuando $k \geq 2$,

$$\phi_k(x) = (x - B_k)\phi_{k-1}(x) - C_k\phi_{k-2}(x), \quad \text{para cada } x \text{ en } [a, b],$$

donde

$$B_k = \frac{\int_a^b x w(x)[\phi_{k-1}(x)]^2\, dx}{\int_a^b w(x)[\phi_{k-1}(x)]^2\, dx}$$

y

$$C_k = \frac{\int_a^b x w(x)\phi_{k-1}(x)\phi_{k-2}(x)\, dx}{\int_a^b w(x)[\phi_{k-2}(x)]^2\, dx}.$$ ■

El teorema 8.7 ofrece un procedimiento recursivo para construir un conjunto de polinomios ortogonales. La demostración de este teorema se logra al aplicar la inducción matemática al grado del polinomio $\phi_n(x)$.

Corolario 8.8 Para toda $n > 0$, el conjunto de las funciones polinomiales $\{\phi_0, \ldots, \phi_n\}$ dado en el teorema 8.7 es linealmente independiente en $[a, b]$ y

$$\int_a^b w(x)\phi_n(x)\, Q_k(x)\, dx = 0,$$

para todo polinomio $Q_k(x)$ de grado $k < n$. ■

Demostración Puesto que $\phi_n(x)$ es un polinomio de n-ésimo grado, el teorema 8.2 implica que $\{\phi_0, \ldots, \phi_n\}$ es un conjunto linealmente independiente.

Sea $Q_k(x)$ un polinomio de k-ésimo grado. De acuerdo con el teorema 8.3, existen los números $c_0, \ldots, c_k$ tales que

$$Q_k(x) = \sum_{j=0}^{k} c_j\, \phi_j(x).$$

Por tanto,

$$\int_a^b w(x)\, Q_k(x)\phi_n(x)\, dx = \sum_{j=0}^{k} c_j \int_a^b w(x)\phi_j(x)\phi_n(x)\, dx = \sum_{j=0}^{k} c_j \cdot 0 = 0,$$

porque ϕ_n es ortogonal a ϕ_j para cada $j = 0, 1, \ldots, k$. ■ ■ ■

EJEMPLO 3 El conjunto de los **polinomios de Legendre**, $\{P_n(x)\}$, es ortogonal en $[-1, 1]$ respecto a la función de peso $w(x) = 1$. La definición clásica de estos polinomios requiere que $P_n(1) = 1$ para cualquier n, y se utiliza una relación recursiva para generar los polinomios cuando $n \geq 2$. No hace falta esta normalización en nuestra exposición; los polinomios de aproximación de mínimos cuadrados que se generan en ambos casos son esencialmente los mismos.

Al aplicar el procedimiento recursivo del teorema 8.7 con $P_0(x) = 1$ obtenemos

$$B_1 = \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} dx} = 0 \quad \text{y} \quad P_1(x) = (x - B_1)P_0(x) = x.$$

También,

$$B_2 = \frac{\int_{-1}^{1} x^3\,dx}{\int_{-1}^{1} x^2\,dx} = 0 \quad \text{y} \quad C_2 = \frac{\int_{-1}^{1} x^2\,dx}{\int_{-1}^{1} 1\,dx} = \frac{1}{3},$$

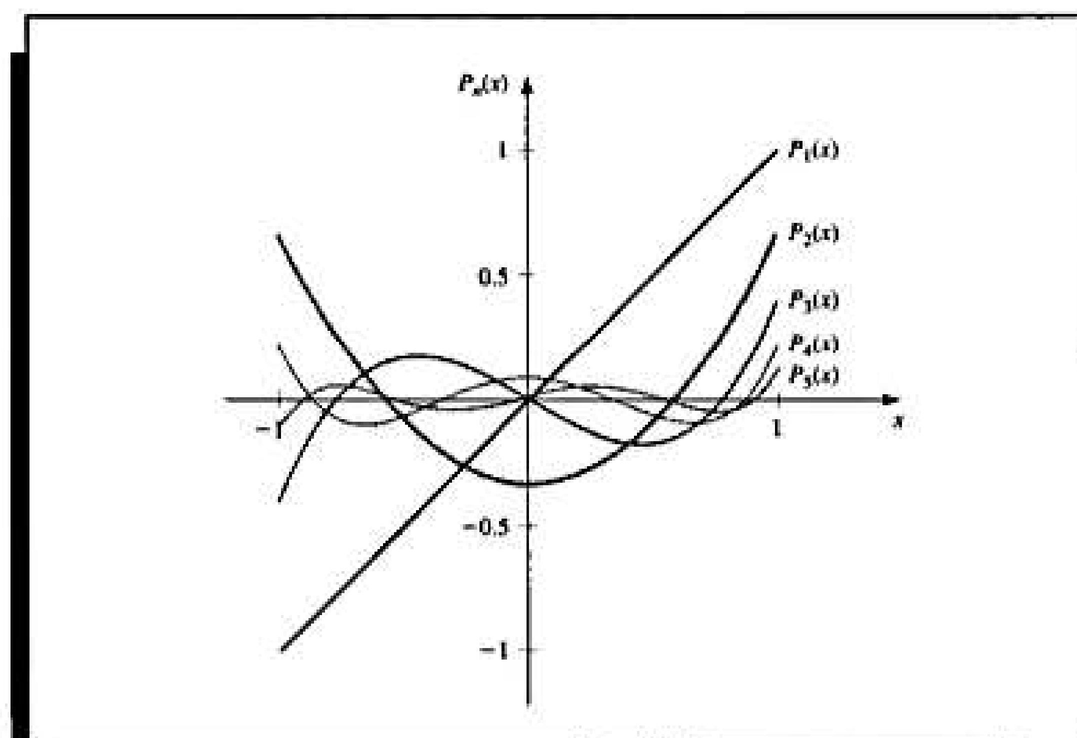
y, por tanto,

$$P_2(x) = (x - B_2)\,P_1(x) - C_2\,P_0(x) = (x - 0)x - \frac{1}{3} \cdot 1 = x^2 - \frac{1}{3}.$$

Los polinomios de Legendre de orden superior que aparecen en la figura 8.9 se deducen de manera análoga. Aunque la integración puede ser tediosa, no es difícil con un sistema de álgebra por computadora. Por ejemplo, el comando de Maple `int` se usa para calcular las integrales B_3 y C_3:

```
>B3:=int(x*(x^2-1/3)^2,x=-1..1)/int(x^2-1/3)^2,x=-1..1);
>C3:=int(x*(x^2-1/3)*x,x=-1..1)/int(x^2,x=-1..1);
```

Figura 8.9



de donde $B_3 = 0$ and $C_3 = \frac{4}{15}$. Así,

$$P_3(x) = x\,P_2(x) - \frac{4}{15}\,P_1(x) = x^3 - \frac{1}{3}x - \frac{4}{15}x = x^3 - \frac{3}{5}x.$$

Los dos siguientes polinomios de Legendre son $P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}$ y $P_5(x) = x^5 - \frac{10}{9}x^3 + \frac{5}{21}x$. ■

Los polinomios de Legendre se mencionaron en la sección 4.7, donde utilizamos sus raíces como los nodos de la cuadratura gaussiana.

CONJUNTO DE EJERCICIOS 8.2

1. Obtenga la aproximación polinomial lineal de mínimos cuadrados a $f(x)$ en el intervalo indicado si
   a. $f(x) = x^2 + 3x + 2$, $[0, 1]$;
   b. $f(x) = x^3$, $[0, 2]$;
   c. $f(x) = \dfrac{1}{x}$, $[1, 3]$;
   d. $f(x) = e^x$, $[0, 2]$;
   e. $f(x) = \dfrac{1}{2}\cos x + \dfrac{1}{3}\operatorname{sen} 2x$, $[0, 1]$;
   f. $f(x) = x \ln x$, $[1, 3]$.
2. Obtenga la aproximación polinomial de mínimos cuadrados de segundo grado a las funciones e intervalos del ejercicio 1.
3. Obtenga la aproximación polinomial lineal en el intervalo $[-1, 1]$ para las siguientes funciones.
   a. $f(x) = x^2 - 2x + 3$
   b. $f(x) = x^3$
   c. $f(x) = \dfrac{1}{x+2}$
   d. $f(x) = e^x$
   e. $f(x) = \dfrac{1}{2}\cos x + \dfrac{1}{3}\operatorname{sen} 2x$
   f. $f(x) = \ln(x + 2)$
4. Obtenga la aproximación polinomial de mínimos cuadrados de segundo grado en el intervalo $[-1, 1]$ para las funciones del ejercicio 3.
5. Calcule el error E para las aproximaciones del ejercicio 3.
6. Calcule el error E para las aproximaciones del ejercicio 4.
7. Utilice el proceso de Gram-Schmidt para construir $\phi_0(x)$, $\phi_1(x)$, $\phi_2(x)$ y $\phi_3(x)$ en los siguientes intervalos.
   a. $[0, 1]$
   b. $[0, 2]$
   c. $[1, 3]$
8. Repita el ejercicio 1 usando los resultados del ejercicio 7.
9. Repita el ejercicio 2 usando los resultados del ejercicio 7.
10. Obtenga el polinomio de aproximación de mínimos cuadrados de tercer grado para las funciones del ejercicio 1 usando los resultados del ejercicio 7.
11. Use el procedimiento de Gram-Schmidt para calcular L_1, L_2 y L_3, donde $\{L_0(x), L_1(x), L_2(x), L_3(x)\}$ es un conjunto ortogonal de polinomios en $(0, \infty)$ respecto a las funciones de peso $w(x) = e^{-x}$ y $L_0(x) \equiv 1$. Los polinomios que se obtienen con este procedimiento reciben el nombre de **polinomios de Laguerre**.

12. Use los polinomios de Laguerre calculados en el ejercicio 11 para obtener los polinomios de mínimos cuadrados de primero, segundo y tercer grados en el intervalo $(0, \infty)$ respecto a la función de peso $w(x) = e^{-x}$ para las siguientes funciones:

    a. $f(x) = x^2$ b. $f(x) = e^{-x}$ c. $f(x) = x^3$ d. $f(x) = e^{-2x}$

13. Suponga que $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un conjunto lineal independiente cualquiera en Π_n. Demuestre que para todo elemento $Q \in \Pi_n$ existen constantes únicas $c_0, c_1, \ldots, c_n$, tales que

$$Q(x) = \sum_{k=0}^{n} c_k\phi_k(x).$$

14. Demuestre que, si $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un conjunto ortogonal de funciones en $[a, b]$ respecto a la función de peso w, entonces $\{\phi_0, \phi_1, \ldots, \phi_n\}$ es un conjunto linealmente independiente.

15. Demuestre que las ecuaciones normales (8.6) tienen una solución única. [*Sugerencia*: demuestre que la única solución de la función $f(x) = 0$ es $a_j = 0, j = 0, 1, \ldots, n$. Multiplique la ecuación (8.6) por a_j y sume sobre todas las j. Intercambie el signo de integral y sumatoria para obtener $\int_a^b [P(x)]^2\,dx = 0$. Por tanto, $P(x) = 0$, así que $a_j = 0$ para $j = 0, \ldots, n$. De aquí que, la matriz de coeficientes es no singular y la ecuación (8.6) tiene una solución única.]

8.3 Polinomios de Chebyshev y economización de las series de potencias

Los polinomios de Chebyshev $\{T_n(x)\}$ son ortogonales en $(-1, 1)$ respecto a la función de peso $w(x) = (1 - x^2)^{-1/2}$. Aunque pueden derivarse por el método que se explicó en la sección anterior, resulta más fácil dar su definición y demostrar después que satisfacen las propiedades de ortogonalidad requeridas.

Con $x \in [-1, 1]$, defina

$$T_n(x) = \cos[n \arccos x], \quad \text{para cada } n \geq 0. \tag{8.8}$$

No es obvio de esta definición que para cada n, $T_n(x)$ es un polinomio en x, pero ahora mostraremos este hecho. Observe primero que

$$T_0(x) = \cos 0 = 1 \quad \text{y} \quad T_1(x) = \cos(\arccos x) = x.$$

Para $n \geq 1$, al introducir la sustitución $\theta = \arccos x$ esta ecuación se transforma en

$$T_n(\theta(x)) \equiv T_n(\theta) = \cos(n\theta), \quad \text{donde } \theta \in [0, \pi].$$

Se deriva una relación de recurrencia al observar que

$$T_{n+1}(\theta) = \cos(n\theta)\cos\theta - \operatorname{sen}(n\theta)\operatorname{sen}\theta$$

y que

$$T_{n-1}(\theta) = \cos(n\theta)\cos\theta + \operatorname{sen}(n\theta)\operatorname{sen}\theta.$$

Al sumar estas ecuaciones obtenemos

$$T_{n+1}(\theta) = 2\cos(n\theta)\cos\theta - T_{n-1}(\theta).$$

Al volver a la variable x, obtenemos, para $n \geq 1$

$$T_{n+1}(x) = 2x \cos(n \arccos x) - T_{n-1}(x)$$

o

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{8.9}$$

Como $T_0(x) = 1$ y $T_1(x) = x$, la relación de recurrencia implica que $T_n(x)$ es un polinomio de grado n con coeficiente principal 2^{n-1}, cuando $n \geq 1$.

Los tres polinomios de Chebyshev que le siguen son

$$T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1,$$
$$T_3(x) = 2xT_2(x) - T_1(x) = 4x^3 - 3x,$$

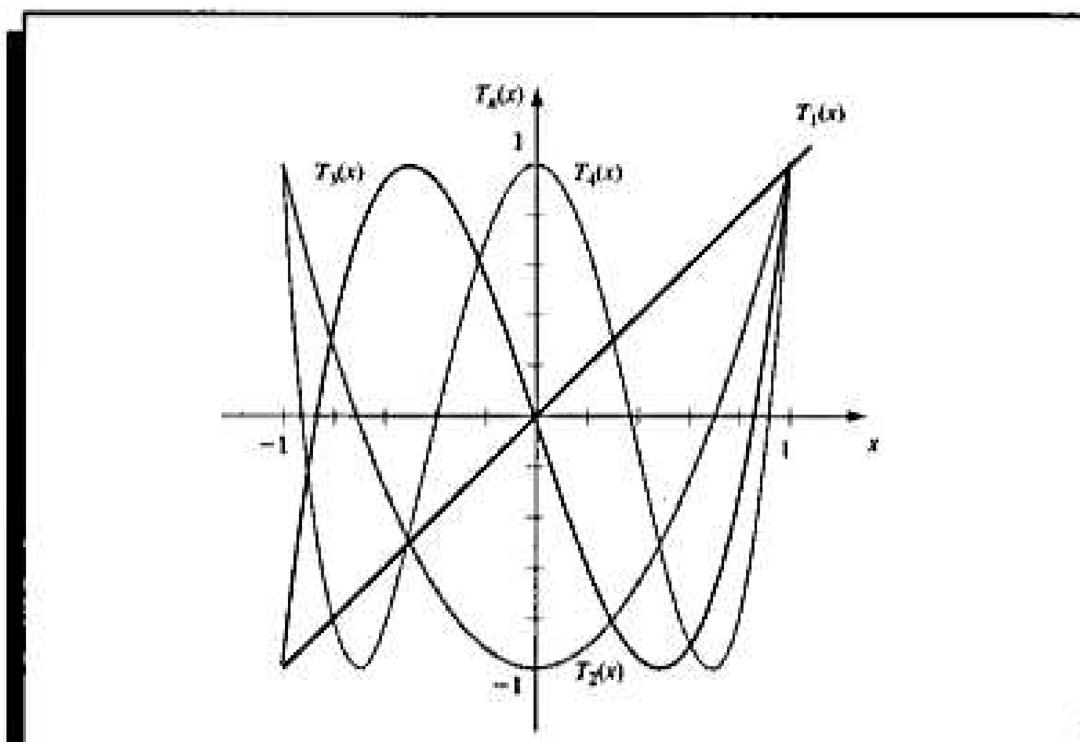
y

$$T_4(x) = 2xT_3(x) - T_2(x) = 8x^4 - 8x^2 + 1.$$

En la figura 8.10 se muestran las gráficas de T_1, T_2, T_3 y T_4. Para demostrar la ortogonalidad de los polinomios de Chebyshev, consideremos

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\, dx = \int_{-1}^{1} \frac{\cos(n \arccos x)\cos(m \arccos x)}{\sqrt{1-x^2}}\, dx.$$

Figura 8.10



Al reintroducir la sustitución $\theta = \arccos x$ se obtiene

$$d\theta = -\frac{1}{\sqrt{1-x^2}}\,dx$$

y

$$\int_{-1}^{1} \frac{T_n(x)\,T_m(x)}{\sqrt{1-x^2}}\,dx = -\int_{\pi}^{0} \cos(n\theta)\cos(m\theta)\,d\theta = \int_{0}^{\pi} \cos(n\theta)\cos(m\theta)\,d\theta.$$

Supongamos que $n \neq m$. Dado que

$$\cos(n\theta)\cos(m\theta) = \frac{1}{2}[\cos(n+m)\theta + \cos(n-m)\theta],$$

tenemos

$$\begin{aligned}\int_{-1}^{1} \frac{T_n(x)\,T_m(x)}{\sqrt{1-x^2}}\,dx &= \frac{1}{2}\int_{0}^{\pi} \cos((n+m)\theta)\,d\theta + \frac{1}{2}\int_{0}^{\pi} \cos((n-m)\theta)\,d\theta \\ &= \left[\frac{1}{2(n+m)}\operatorname{sen}((n+m)\theta) + \frac{1}{2(n-m)}\operatorname{sen}((n-m)\theta)\right]_{0}^{\pi} \\ &= 0.\end{aligned}$$

Usando un método similar también podemos demostrar que, cuando $n = m$,

$$\int_{-1}^{1} \frac{[T_n(x)]^2}{\sqrt{1-x^2}}\,dx = \frac{\pi}{2}, \quad \text{para cada } n \geq 1. \tag{8.10}$$

Los polinomios de Chebyshev sirven para reducir al mínimo el error de aproximación. Veremos cómo se usan para resolver dos problemas de este tipo:

1. Una colocación óptima de los puntos interpolantes para reducir al mínimo el error en la interpolación de Lagrange.
2. Un medio de reducir el grado de un polinomio de aproximación con una pérdida de exactitud mínima.

El siguiente resultado se refiere a los ceros y a los puntos extremos de T_n.

Teorema 8.9 El polinomio de Chebyshev $T_n(x)$ de grado $n \geq 1$ tiene n ceros simples en $[-1, 1]$ en

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{para cada } k = 1, 2, \ldots, n.$$

Más aún, $T_n(x)$ toma sus extremos absolutos en

$$\bar{x}'_k = \cos\left(\frac{k\pi}{n}\right), \quad \text{con} \quad T_n(\bar{x}'_k) = (-1)^k, \quad \text{para cada } k = 0, 1, \ldots, n. \quad \blacksquare$$

Demostración Si utilizamos

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{para cada } k = 1, 2, \ldots, n,$$

entonces

$$T_n(\bar{x}_k) = \cos(n \arccos \bar{x}_k) = \cos\left(n \arccos\left(\cos\left(\frac{2k-1}{2n}\pi\right)\right)\right) = \cos\left(\frac{2k-1}{2}\pi\right) = 0,$$

y cada $\bar{x}_k$ es un cero distinto de T_n. Por ser $T_n(x)$ un polinomio de n-ésimo grado, todos los ceros de T_n deben ser de esta misma forma.

Para demostrar la segunda parte, primero obsérvese que

$$T'_n(x) = \frac{d}{dx}[\cos(n \arccos x)] = \frac{n \operatorname{sen}(n \arccos x)}{\sqrt{1-x^2}},$$

y que, cuando $k = 1, 2, \ldots, n-1$,

$$T'_n(\bar{x}'_k) = \frac{n \operatorname{sen}\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right)}{\sqrt{1-\left[\cos\left(\frac{k\pi}{n}\right)\right]^2}} = \frac{n \operatorname{sen}(k\pi)}{\operatorname{sen}\left(\frac{k\pi}{n}\right)} = 0.$$

Como $T_n(x)$ es un polinomio de n-ésimo grado, su derivada $T'_n(x)$ es un polinomio de $(n-1)$-ésimo grado y todos los ceros de $T'_n(x)$ ocurren en estos $n-1$ puntos. Las otras únicas posibilidades de los extremos de $T_n(x)$ se presentan en los extremos del intervalo $[-1, 1]$; es decir, $\bar{x}'_0 = 1$ y $\bar{x}'_n = -1$. Por tanto, para cualquier $k = 0, 1, \ldots, n$, tenemos

$$T_n(\bar{x}'_k) = \cos\left(n \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right) = \cos(k\pi) = (-1)^k,$$

en cada valor par de k ocurre un máximo, y en cada valor impar un mínimo. ■ ■ ■

Los polinomios mónicos de Chebyshev (los que tienen coeficiente principal 1) $\tilde{T}_n(x)$ se derivan del polinomio de Chebyshev $T_n(x)$ al dividirlos por el coeficiente principal 2^{n-1}. Por tanto,

$$\tilde{T}_0(x) = 1 \quad \text{y} \quad \tilde{T}_n(x) = \frac{1}{2^{n-1}} T_n(x), \quad \text{para cada } n \geq 1. \tag{8.11}$$

La relación de recurrencia satisfecha por los polinomios de Chebyshev implica que

$$\tilde{T}_2(x) = x\,\tilde{T}_1(x) - \frac{1}{2}\tilde{T}_0(x) \quad \text{y} \tag{8.12}$$

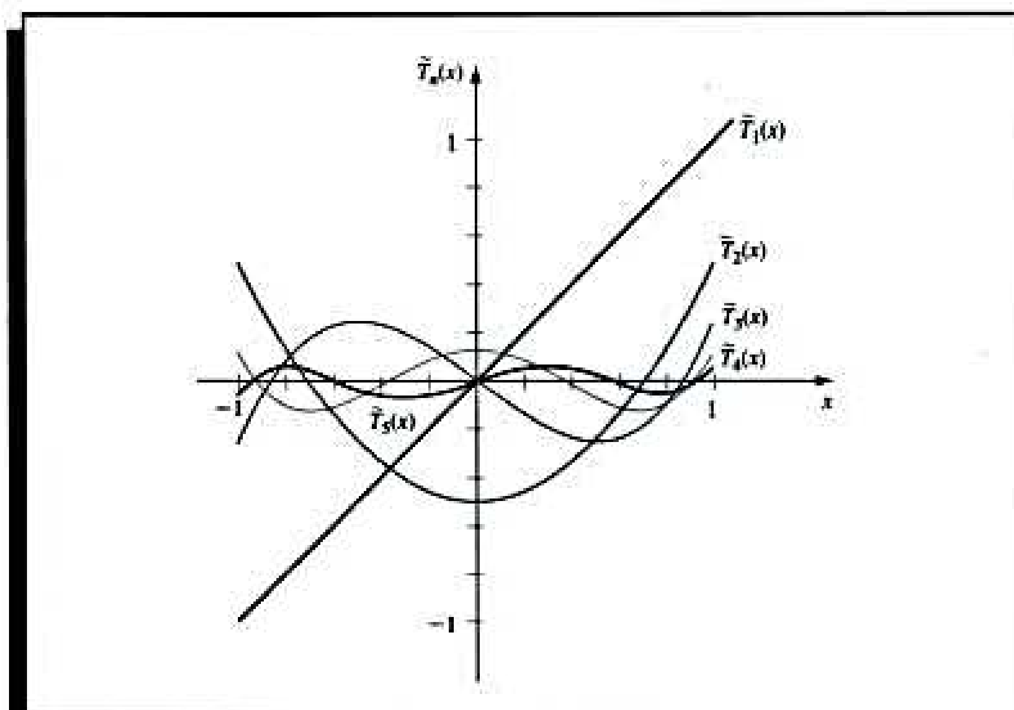
$$\tilde{T}_{n+1}(x) = x\,\tilde{T}_n(x) - \frac{1}{4}\tilde{T}_{n-1}(x), \quad \text{para cada } n \geq 2.$$

En la figura 8.11 se muestran las gráficas de $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{T}_3$, $\tilde{T}_4$ y $\tilde{T}_5$.

Como $\tilde{T}_n(x)$ es sólo un múltiplo de $T_n(x)$, el teorema 8.9 implica que los ceros de $\tilde{T}_n(x)$ también ocurren en

$$\bar{x}_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad \text{para cada } k = 1, 2, \ldots, n.$$

Figura 8.11



y que los valores extremos de $\tilde{T}_n(x)$, para $n \geq 1$, se presentan en

$$\bar{x}'_k = \cos\left(\frac{k\pi}{n}\right), \quad \text{con} \quad \tilde{T}_n(\bar{x}'_k) = \frac{(-1)^k}{2^{n-1}}, \quad \text{para cada } k = 0, 1, 2, \ldots, n. \quad (8.13)$$

Denotemos con $\tilde{\prod}_n$ **el conjunto de todos los polinomios mónicos de n-ésimo grado.** En la ecuación (8.13) se expresa la relación que da origen a una importante propiedad de minimización que distingue a $\tilde{T}_n(x)$ de otros miembros de $\tilde{\prod}_n$.

Teorema 8.10 Los polinomios de la forma $\tilde{T}_n(x)$, cuando $n \geq 1$, tienen la propiedad de que

$$\frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)| \leq \max_{x\in[-1,1]} |P_n(x)|, \quad \text{para cualquier } P_n(x) \in \tilde{\prod}_n.$$

Además, la igualdad puede darse sólo si $P_n = \tilde{T}_n$. ■

Demostración Supongamos que $P_n(x) \in \tilde{\prod}_n$ y que

$$\max_{x\in[-1,1]} |P_n(x)| \leq \frac{1}{2^{n-1}} = \max_{x\in[-1,1]} |\tilde{T}_n(x)|.$$

Sea $Q = \tilde{T}_n - P_n$. Como $\tilde{T}_n(x)$ y $P_n(x)$ son polinomios mónicos de n-ésimo grado, $Q(x)$ es un polinomio de grado a lo más $(n-1)$. Más aún, en los puntos extremos de $\tilde{T}_n(x)$,

$$Q(\bar{x}'_k) = \tilde{T}_n(\bar{x}'_k) - P_n(\bar{x}'_k) = \frac{(-1)^k}{2^{n-1}} - P_n(\bar{x}'_k).$$

Dado que

$$|P_n(\bar{x}'_k)| \leq \frac{1}{2^{n-1}}, \quad \text{para cada } k = 0, 1, \ldots, n,$$

tenemos

$$Q(\bar{x}'_k) \leq 0 \quad \text{cuando } k \text{ es impar} \quad \text{y} \quad Q(\bar{x}'_k) \geq 0 \quad \text{cuando } k \text{ es par.}$$

Como Q es continua, el teorema de valor intermedio implica que el polinomio $Q(x)$ tiene, al menos, un cero entre $\bar{x}'_j$ y $\bar{x}'_{j+1}$ para cada $j = 0, 1, \ldots, n-1$. Por tanto, Q tiene al menos n ceros en el intervalo $[-1, 1]$. Pero el grado de $Q(x)$ es menor que n, así que $Q \equiv 0$. Ello implica que $P_n \equiv \tilde{T}_n$. ■ ■ ■

Este teorema sirve para contestar la pregunta de dónde colocar los nodos interpolantes para reducir al mínimo el error de la interpolación de Lagrange. El teorema 3.3 aplicado al intervalo $[-1, 1]$ establece que si, $x_0, \ldots, x_n$ son números distintos en el intervalo $[-1, 1]$ y si $f \in C^{n+1}[-1, 1]$, entonces, para cada $x \in [-1, 1]$, existe un número $\xi(x)$ en $(-1, 1)$ con

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n),$$

donde $P(x)$ denota el polinomio interpolante de Lagrange. En general, no se tiene control sobre $\xi(x)$ por lo cual, para reducir al mínimo el error mediante una colocación habilidosa de los nodos $x_0, \ldots, x_n$, obtenemos $x_0, \ldots, x_n$ para reducir al mínimo la cantidad

$$|(x - x_0)(x - x_1)\cdots(x - x_n)|$$

a lo largo del intervalo $[-1, 1]$.

Puesto que $(x - x_0)(x - x_1)\cdots(x - x_n)$ es un polinomio mónico de grado $(n+1)$, acabamos de ver que el mínimo se obtiene si y sólo si

$$(x - x_0)(x - x_1)\cdots(x - x_n) = \tilde{T}_{n+1}(x).$$

El valor máximo de $|(x - x_0)(x - x_1)\cdots(x - x_n)|$ se minimiza cuando hacemos que x_k sea el $(k+1)$-ésimo cero de $\tilde{T}_{n+1}$, para cada $k = 0, 1, \ldots, n$, esto es, cuando x_k es

$$\bar{x}_{k+1} = \cos\frac{2k+1}{2(n+1)}\pi.$$

Como $\max_{x\in[-1,1]}|\tilde{T}_{n+1}(x)| = 2^{-n}$, esto también implica que

$$\frac{1}{2^n} = \max_{x\in[-1,1]}|(x - \bar{x}_1)\cdots(x - \bar{x}_{n+1})| \leq \max_{x\in[-1,1]}|(x - x_0)\cdots(x - x_n)|$$

para cualquier selección de $x_0, x_1, \ldots, x_n$ en el intervalo $[-1, 1]$. El siguiente corolario se extrae de esta explicación.

Corolario 8.11 Si $P(x)$ es el polinomio interpolante de grado a lo más n en las raíces de $T_{n+1}(x)$, entonces

$$\max_{x\in[-1,1]}|f(x) - P(x)| \leq \frac{1}{2^n(n+1)!}\max_{x\in[-1,1]}|f^{(n+1)}(x)|, \quad \text{para cada } f \in C^{n+1}[-1, 1].$$

■

Este método de seleccionar puntos para reducir al mínimo el error de interpolación, también se aplica a un intervalo cerrado general $[a, b]$ empleando el cambio de variables

$$\tilde{x} = \frac{1}{2}[(b - a)x + a + b]$$

para transformar los números $\bar{x}_k$ del intervalo $[-1, 1]$ en el correspondiente número $\tilde{x}_k$ del intervalo $[a, b]$ como se muestra en el siguiente ejemplo.

EJEMPLO 1 Sea $f(x) = xe^x$ en el intervalo $[0, 1.5]$. Se construirán dos polinomios interpolantes de grado tres como máximo. Primero, empleamos los nodos uniformemente espaciados $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$ y $x_3 = 1.5$ para obtener

$$L_0(x) = -1.3333x^3 + 4.0000x^2 - 3.6667x + 1,$$
$$L_1(x) = 4.0000x^3 - 10.000x^2 + 6.0000x,$$
$$L_2(x) = -4.0000x^3 + 8.0000x^2 - 3.0000x,$$
$$L_3(x) = 1.3333x^3 - 2.000x^2 + 0.66667x.$$

En el caso de los valores que se dan en las dos primeras columnas de la tabla 8.7, el primer polinomio está dado por

$$P_3(x) = 1.3875x^3 + 0.057570x^2 + 1.2730x.$$

Tabla 8.7

x	$f(x) = xe^x$	$\tilde{x}$	$f(\tilde{x}) = xe^x$
$x_0 = 0.0$	0.00000	$\tilde{x}_0 = 1.44291$	6.10783
$x_1 = 0.5$	0.824361	$\tilde{x}_1 = 1.03701$	2.92517
$x_2 = 1.0$	2.71828	$\tilde{x}_2 = 0.46299$	0.73560
$x_3 = 1.5$	6.72253	$\tilde{x}_3 = 0.05709$	0.060444

En el caso del segundo polinomio interpolante, se transforman los ceros $\bar{x}_k = \cos((2k + 1)/8)\pi$, cuando $k = 0, 1, 2, 3$, de $\tilde{T}_4$, desde $[-1, 1]$ a $[0, 1.5]$ mediante la transformación lineal

$$\tilde{x}_k = \frac{1}{2}[(1.5 - 0)\,\bar{x}_k + (1.5 + 0)] = 0.75 + 0.75\,\bar{x}_k$$

para obtener

$$\tilde{x}_0 = 1.44291, \quad \tilde{x}_1 = 1.03701, \quad \tilde{x}_2 = 0.46299 \quad \text{y} \quad \tilde{x}_3 = 0.05709.$$

Entonces, para este conjunto de nodos, los coeficientes polinomiales de Lagrange se calculan así:

$$\tilde{L}_0(x) = 1.8142x^3 - 2.8249x^2 + 1.0264x - 0.049728,$$
$$\tilde{L}_1(x) = -4.3799x^3 + 8.5977x^2 - 3.4026x + 0.16705,$$
$$\tilde{L}_2(x) = 4.3799x^3 - 11.112x^2 + 7.1738x - 0.37415,$$
$$\tilde{L}_3(x) = -1.8142x^3 + 5.3390x^2 - 4.7976x + 1.2568.$$

Los valores funcionales que se requieren en estos polinomios se dan en las dos últimas columnas de la tabla 8.7. El polinomio interpolante de grado a lo más tres está dado por

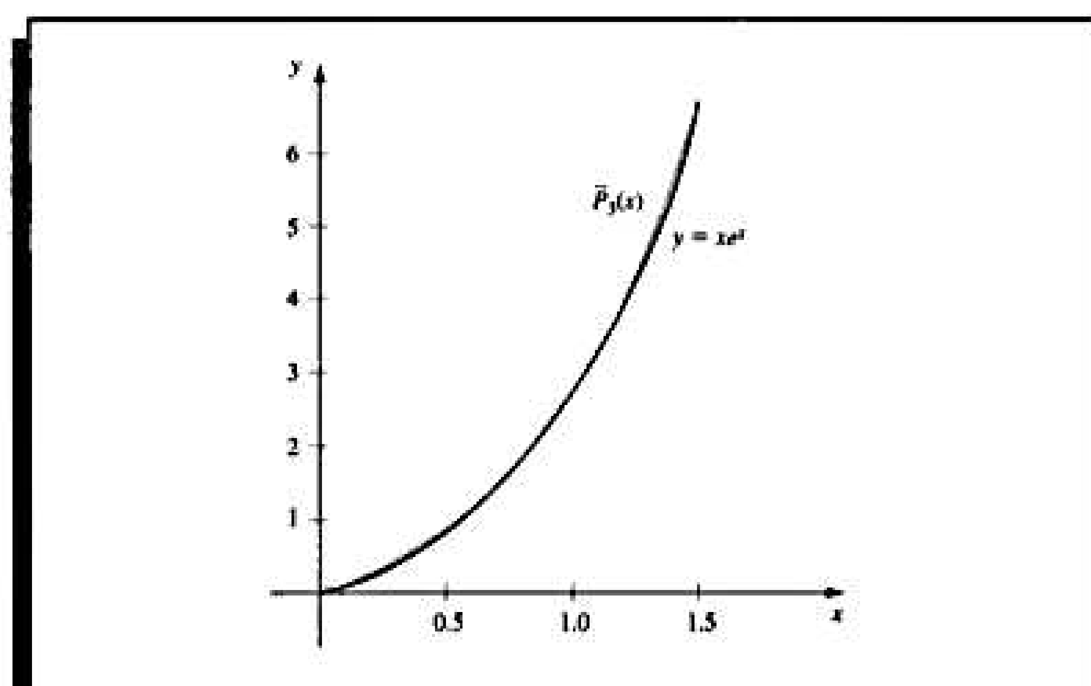
$$\tilde{P}_3(x) = 1.3811x^3 + 0.044652x^2 + 1.3031x - 0.014352.$$

Para efectos de comparación, en la tabla 8.8 se incluyen varios valores de x, junto con los valores de $f(x)$, $P_3(x)$ y de $\tilde{P}_3(x)$. De la tabla se puede ver que, aunque el error al utilizar $P_3(x)$ es menor que al utilizar $\tilde{P}_3(x)$ cerca de la mitad de la tabla, el error máximo que implica usar $\tilde{P}_3(x)$, 0.0180, es considerablemente menor que cuando se emplea $P_3(x)$, lo cual da el error 0.0290. (Véase la Fig. 8.12.) ■

Tabla 8.8

x	$f(x) = xe^x$	$P_3(x)$	$\lvert xe^x - P_3(x)\rvert$	$\tilde{P}_3(x)$	$\lvert xe^x - \tilde{P}_3(x)\rvert$
0.15	0.1743	0.1969	0.0226	0.1868	0.0125
0.25	0.3210	0.3435	0.0225	0.3358	0.0148
0.35	0.4967	0.5121	0.0154	0.5064	0.0097
0.65	1.245	1.233	0.012	1.231	0.014
0.75	1.588	1.572	0.016	1.571	0.017
0.85	1.989	1.976	0.013	1.974	0.015
1.15	3.632	3.650	0.018	3.644	0.012
1.25	4.363	4.391	0.028	4.382	0.019
1.35	5.208	5.237	0.029	5.224	0.016

Figura 8.12



Los polinomios de Chebyshev también sirven para disminuir el grado de un polinomio de aproximación, con una pérdida mínima de exactitud. Como estos polinomios tienen un valor absoluto mínimo/máximo que se distribuye uniformemente en el intervalo,

pueden usarse para reducir el grado de un polinomio de aproximación, sin que se rebase la tolerancia de error.

Por ejemplo, aproximemos un polinomio arbitrario de n-ésimo grado

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

en $[-1, 1]$ con un polinomio de grado $n-1$ como máximo. El objetivo es seleccionar $P_{n-1}(x)$ en $\prod_{n-1}$ de modo que

$$\max_{x\in[-1,1]} |P_n(x) - P_{n-1}(x)|$$

sea lo menor posible.

Primero observamos que $(P_n(x) - P_{n-1}(x))/a_n$ es un polinomio mónico de n-ésimo grado. Al aplicar el teorema 8.10 obtenemos

$$\max_{x\in[-1,1]} \left|\frac{1}{a_n}(P_n(x) - P_{n-1}(x))\right| \geq \frac{1}{2^{n-1}}.$$

La igualdad ocurre precisamente cuando

$$\frac{1}{a_n}(P_n(x) - P_{n-1}(x)) = \tilde{T}_n(x).$$

Esto significa que deberemos seleccionar

$$P_{n-1}(x) = P_n(x) - a_n \tilde{T}_n(x),$$

y con esta elección tenemos el valor mínimo de

$$\max_{x\in[-1,1]} |P_n(x) - P_{n-1}(x)| = |a_n| \max_{x\in[-1,1]} \left|\frac{1}{a_n}(P_n(x) - P_{n-1}(x))\right| = \frac{|a_n|}{2^{n-1}}.$$

EJEMPLO 2 Aproximaremos la función $f(x) = e^x$ en el intervalo $[-1, 1]$ mediante el cuarto polinomio de Maclaurin

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24},$$

que tiene el error de truncamiento

$$|R_4(x)| = \frac{|f^{(5)}(\xi(x))||x^5|}{120} \leq \frac{e}{120} \approx 0.023, \quad \text{para } -1 \leq x \leq 1.$$

Supongamos que un error de 0.05 es tolerable y que nos gustaría reducir el grado del polinomio de aproximación sin rebasar esta cota.

El polinomio de tercer grado, o menor, que mejor aproxima uniformemente $P_4(x)$ en $[-1, 1]$ es

$$P_3(x) = P_4(x) - a_4\tilde{T}_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} - \frac{1}{24}\left(x^4 - x^2 + \frac{1}{8}\right)$$
$$= \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3.$$

Con esta selección tenemos

$$|P_4(x) - P_3(x)| = |a_4 \tilde{T}_4(x)| \leq \frac{1}{24} \cdot \frac{1}{2^3} = \frac{1}{192} \leq 0.0053.$$

Al agregar esta cota de error a la cota del error de truncamiento de Maclaurin, obtenemos

$$0.023 + 0.0053 = 0.0283,$$

que se encuentra dentro del error permisible de 0.05.

El polinomio de segundo o menor grado que mejor aproxima uniformemente a $P_3(x)$ en $[-1, 1]$ es

$$P_2(x) = P_3(x) - \frac{1}{6} \tilde{T}_3(x)$$

$$= \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3 - \frac{1}{6}\left(x^3 - \frac{3}{4}x\right) = \frac{191}{192} + \frac{9}{8}x + \frac{13}{24}x^2.$$

Sin embargo,

$$|P_3(x) - P_2(x)| = \left|\frac{1}{6}\tilde{T}_3(x)\right| = \frac{1}{6}\left(\frac{1}{2}\right)^2 = \frac{1}{24} \approx 0.042,$$

que, al ser agregado a la cota de error ya acumulada de 0.0283, rebasa la tolerancia de 0.05. En consecuencia, el polinomio de menor grado que mejor aproxima a e^x en $[-1, 1]$ con una cota de error menor que 0.05 es

$$P_3(x) = \frac{191}{192} + x + \frac{13}{24}x^2 + \frac{1}{6}x^3.$$

La tabla 8.9 incluye la función y los polinomios de aproximación en varios puntos de $[-1, 1]$. Nótese que los valores tabulados de P_2 se encuentran, con mucho, dentro de la tolerancia de 0.05, aunque la cota de error de $P_2(x)$ la rebasa. ■

Tabla 8.9

x	e^x	$P_4(x)$	$P_3(x)$	$P_2(x)$	$\lvert e^x - P_2(x)\rvert$
−0.75	0.47237	0.47412	0.47917	0.45573	0.01664
−0.25	0.77880	0.77881	0.77604	0.74740	0.03140
0.00	1.00000	1.00000	0.99479	0.99479	0.00521
0.25	1.28403	1.28402	1.28125	1.30990	0.02587
0.75	2.11700	2.11475	2.11979	2.14323	0.02623

CONJUNTO DE EJERCICIOS 8.3

1. Use los ceros de $\tilde{T}_3$ y construya un polinomio interpolante de segundo grado para las siguientes funciones en el intervalo $[-1, 1]$.

   a. $f(x) = e^x$
   b. $f(x) = \operatorname{sen} x$
   c. $f(x) = \ln(x + 2)$
   d. $f(x) = x^4$

2. Encuentre las cotas para el error máximo de las aproximaciones del ejercicio 1 en el intervalo $[-1, 1]$.
3. Use los ceros de $\tilde{T}_4$ y construya un polinomio interpolante de tercer grado para las funciones del ejercicio 1.
4. Repita el ejercicio 2 con las aproximaciones calculadas en el ejercicio 3.
5. Use los ceros de $\tilde{T}_3$ y las transformaciones del intervalo dado, y construya un polinomio interpolante de segundo grado para las siguientes funciones.

   a. $f(x) = \dfrac{1}{x}$, $[1, 3]$
   b. $f(x) = e^{-x}$, $[0, 2]$
   c. $f(x) = \dfrac{1}{2}\cos x + \dfrac{1}{3}\operatorname{sen} 2x$, $[0, 1]$
   d. $f(x) = x \ln x$, $[1, 3]$

6. Encuentre el sexto polinomio de Maclaurin para xe^x, y con la economización de Chebyshev obtenga una aproximación de menor grado, sin que el error sea mayor que 0.01 en $[-1, 1]$.
7. Encuentre el sexto polinomio de Maclaurin para sen x y use la economización de Chebyshev para obtener una aproximación polinomial de menor grado, sin que el error sea mayor que 0.01 en $[-1, 1]$.
8. Demuestre que para cualesquiera enteros positivos i y j con $i > j$, tenemos que $T_i(x)T_j(x) = \frac{1}{2}[T_{i+j}(x) + T_{i-j}(x)]$.
9. Demuestre que para cualquier polinomio de Chebyshev $T_n(x)$, tenemos

$$\int_{-1}^{1} \frac{[T_n(x)]^2}{\sqrt{1-x^2}}\,dx = \frac{\pi}{2}.$$

8.4 Aproximación mediante la función racional

Los polinomios algebraicos ofrecen ventajas muy claras que pueden emplearse en la aproximación:

1. se cuenta con suficientes polinomios para aproximar una función continua en un intervalo cerrado sin exceder una tolerancia arbitraria;
2. los polinomios se evalúan fácilmente en valores arbitrarios y
3. hay derivadas e integrales de polinomios que pueden determinarse fácilmente.

La desventaja del uso de los polinomios en la aproximación es su tendencia a oscilar. A menudo esto hace que en la aproximación polinomial las cotas de error rebasen significativamente el error promedio de aproximación, pues las cotas de error se determinan mediante el error máximo de aproximación. En esta sección estudiaremos los métodos que distribuyen ese error más uniformemente en el intervalo de aproximación. En estas técnicas intervienen las funciones racionales.

Una **función racional** r de grado N tiene la forma

$$r(x) = \frac{p(x)}{q(x)},$$

donde $p(x)$ y $q(x)$ son los polinomios cuyos grados suman N.

Todo polinomio es una función racional (tomemos simplemente $q(x) \equiv 1$), y por eso la aproximación mediante funciones racionales genera resultados que no son peores a los que se consiguen por medio de la aproximación con polinomios. Sin embargo, las funcio-

nes racionales cuyo numerador y denominador tienen el mismo grado o un grado casi idéntico producen resultados superiores a los métodos que usan polinomios, y ello con la misma cantidad de cálculos. (Esta afirmación se basa en la suposición de que los cálculos requeridos en la división son aproximadamente iguales a los de la multiplicación.) Las funciones racionales ofrecen una ventaja más: permiten una aproximación eficiente de las funciones que tienen discontinuidades infinitas cerca del intervalo de aproximación, pero fuera de él. En este caso, la aproximación polinomial casi siempre resulta inaceptable.

Supongamos que r es una función racional de grado $N = n + m$ de la forma

$$r(x) = \frac{p(x)}{q(x)} = \frac{p_0 + p_1x + \cdots + p_nx^n}{q_0 + q_1x + \cdots + q_mx^m},$$

por medio de la cual aproximamos una función f en un intervalo cerrado I que contiene al cero. Para que r esté definida en cero se requiere que $q_0 \neq 0$. Podemos suponer que $q_0 = 1$, porque de no ser así simplemente reemplazamos $p(x)$ por $p(x)/q_0$ y $q(x)$ por $q(x)/q_0$. En consecuencia, hay $N + 1$ parámetros $q_1, q_2, \ldots, q_m, p_0, p_1, \ldots, p_n$ disponibles para aproximar f por medio de r.

El **método de aproximación de Padé** selecciona los $N + 1$ parámetros de modo que $f^{(k)}(0) = r^{(k)}(0)$ para cada $k = 0, 1, \ldots, N$. La aproximación de Padé es una extensión de la aproximación polinomial de Taylor a las funciones racionales. De hecho, cuando $n = N$ y $m = 0$, la aproximación de Padé es simplemente el N−ésimo polinomio de Maclaurin.

Consideremos la diferencia

$$f(x) - r(x) = f(x) - \frac{p(x)}{q(x)} = \frac{f(x)\,q(x) - p(x)}{q(x)} = \frac{f(x)\sum_{i=0}^{m} q_ix^i - \sum_{i=0}^{n} p_ix^i}{q(x)},$$

y supongamos que f tiene la expansión de la serie de Maclaurin $f(x) = \sum_{i=0}^{\infty} a_ix^i$. Entonces

$$f(x) - r(x) = \frac{\sum_{i=0}^{\infty} a_ix^i \sum_{i=0}^{m} q_ix^i - \sum_{i=0}^{n} p_ix^i}{q(x)}. \tag{8.14}$$

El objetivo es seleccionar las constantes $q_1, q_2, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$ de tal forma que

$$f^{(k)}(0) - r^{(k)}(0) = 0, \quad \text{para cada } k = 0, 1, \ldots, N.$$

En la sección 2.4 (véase, en especial, el ejercicio 10) comprobamos que lo anterior equivale a que $f - r$ tenga un cero de multiplicidad $N + 1$ en $x = 0$. En consecuencia, seleccionamos $q_1, q_2, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$ de manera que el numerador del lado derecho de la ecuación (8.14),

$$(a_0 + a_1x + \cdots)(1 + q_1x + \cdots + q_mx^m) - (p_0 + p_1x + \cdots + p_nx^n), \tag{8.15}$$

no tenga términos de un grado menor o igual que N.

Con el fin de simplificar la notación definimos $p_{n+1} = p_{n+2} = \cdots = p_N = 0$ y $q_{m+1} = q_{m+2} = \cdots = q_N = 0$. Ahora podemos escribir el coeficiente de x^k en la expresión (8.15) así

$$\left(\sum_{i=0}^{k} a_iq_{k-i}\right) - p_k.$$

Por tanto, la función racional de la aproximación de Padé proviene de la solución de las $N + 1$ ecuaciones lineales:

$$\sum_{i=0}^{k} a_i q_{k-i} = p_k, \quad k = 0, 1, \ldots, N$$

en las $N + 1$ incógnitas $q_1, q_2, \ldots, q_m, p_0, p_1, \ldots, p_n$.

EJEMPLO 1 El desarrollo de la serie de Maclaurin para e^{-x} es

$$\sum_{i=0}^{\infty} \frac{(-1)^i}{i!} x^i.$$

Para encontrar la aproximación de Padé a e^{-x} de quinto grado con $n = 3$ y $m = 2$ se requiere seleccionar p_0, p_1, p_2, p_3, q_1 y q_2 de manera que los coeficientes de x^k para $k = 0, 1, \ldots, 5$ sean cero en la expresión

$$\left(1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots\right)(1 + q_1 x + q_2 x^2) - (p_0 + p_1 x + p_2 x^2 + p_3 x^3).$$

Al expandir y agrupar los términos obtenemos

$$x^5: \quad -\frac{1}{120} + \frac{1}{24} q_1 - \frac{1}{6} q_2 = 0; \qquad x^2: \quad \frac{1}{2} - q_1 + q_2 = p_2;$$

$$x^4: \quad \frac{1}{24} - \frac{1}{6} q_1 + \frac{1}{2} q_2 = 0; \qquad x^1: \quad -1 + q_1 = p_1;$$

$$x^3: \quad -\frac{1}{6} + \frac{1}{2} q_1 - q_2 = p_3; \qquad x^0: \quad 1 = p_0.$$

Para resolver el sistema en Maple usamos los siguientes comandos:

```
>eq1:=-1+q1=p1;
>eq2:=1/2-q1+q2=p2;
>eq3:=-1/6+1/2*q1-q2=p3;
>eq4:=1/24-1/6*q1+1/2*q2=0;
>eq5:=1/120+1/24*q1-1/6*q2=0;
>solve({eq1,eq2,eq3,eq4,eq5},{q1,q2,p1,p2,p3});
```

lo que da

$$p_0 = 1, \quad p_1 = -\frac{3}{5}, p_2 = \frac{3}{20}, \quad p_3 = -\frac{1}{60}, \quad q_1 = \frac{2}{5} \quad \text{y} \quad q_2 = \frac{1}{20}.$$

Tabla 8.10

x	e^{-x}	$P_5(x)$	$\lvert e^{-x} - P_5(x)\rvert$	$r(x)$	$\lvert e^{-x} - r(x)\rvert$
0.2	0.81873075	0.81873067	8.64×10^{-8}	0.81873075	7.55×10^{-9}
0.4	0.67032005	0.67031467	5.38×10^{-6}	0.67031963	4.11×10^{-7}
0.6	0.54881164	0.54875200	5.96×10^{-5}	0.54880763	4.00×10^{-6}
0.8	0.44932896	0.44900267	3.26×10^{-4}	0.44930966	1.93×10^{-5}
1.0	0.36787944	0.36666667	1.21×10^{-3}	0.36781609	6.33×10^{-5}

Por tanto, la aproximación de Padé es

$$r(x) = \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2}.$$

En la tabla 8.10 se dan los valores de $r(x)$ y $P_5(x)$, el quinto polinomio de Maclaurin. En este ejemplo la aproximación de Padé es evidentemente superior. ■

Con Maple puede calcularse la aproximación de Padé. Primero se calcula la serie de Maclaurin con el comando:

```
>series(exp(-x),x);
```

para obtener

$$1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{120}x^5 + O(x^6).$$

La aproximación de Padé con $n = 3$ y $m = 2$ se calcula mediante el comando

```
>g:=convert(%,ratpoly,3,2);
```

donde % se refiere al resultado del cálculo anterior, o sea a la serie. El resultado es

$$g := \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2},$$

y entonces podemos calcular $g(0.8)$ introduciendo

```
>evalf(subs(x=0.8,g));
```

para obtener .4493096647.

En el algoritmo 8.1 se ejecuta el método de aproximación de Padé.

ALGORITMO 8.1

Método de aproximación racional de Padé

Para obtener la aproximación racional

$$r(x) = \frac{p(x)}{q(x)} = \frac{\sum_{i=0}^{n} p_i x^i}{\sum_{j=0}^{m} q_j x^j}$$

para determinada función $f(x)$:

ENTRADA enteros no negativos m y n.

SALIDA coeficientes $q_0, q_1, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$.

Paso 1 Tome $N = m + n$.

Paso 2 Para $i = 0, 1, \ldots, N$ tome $a_i = \dfrac{f^{(i)}(0)}{i!}$.

(Los coeficientes del polinomio de Maclaurin son $a_0, \ldots, a_N$, los cuales se pueden introducir en lugar de calcularse.)

Paso 3 Tome $q_0 = 1$;
$p_0 = a_0$.

Paso 4 Para $i = 1, 2, \ldots, N$ haga los pasos 5-10. *(Utilice un sistema lineal con matriz B.)*

Paso 5 Para $j = 1, 2, \ldots, i - 1$
si $j \leq n$ entonces tome $b_{i,j} = 0$.

Paso 6 Si $i \leq n$ entonces tome $b_{i,i} = 1$.

Paso 7 Para $j = i + 1, i + 2, \ldots, N$ tome $b_{i,j} = 0$.

Paso 8 Para $j = 1, 2, \ldots, i$
si $j \leq m$ entonces tome $b_{i,n+j} = -a_{i-j}$.

Paso 9 Para $j = n + i + 1, n + i + 2, \ldots, N$ tome $b_{i,j} = 0$.

Paso 10 Tome $b_{i,N+1} = a_i$.

(Pasos 11-22 resuelven el sistema lineal usando el pivoteo parcial.)

Paso 11 Para $i = n + 1, n + 2, \ldots, N - 1$ haga los pasos 12-18.

Paso 12 Sea k el entero más pequeño con $i \leq k \leq N$ y $|b_{k,i}| = \text{máx}_{i\leq j\leq N} |b_{j,i}|$. *(Obtenga el elemento de pivote.)*

Paso 13 Si $b_{k,i} = 0$ entonces SALIDA ("El sistema es singular");
PARAR.

Paso 14 Si $k \neq i$ entonces *(Intercambie el renglón i y el renglón k.)*
para $j = i, i + 1, \ldots, N + 1$ tome

$$b_{COPY} = b_{i,j};$$
$$b_{i,j} = b_{k,j};$$
$$b_{k,j} = b_{COPY}.$$

Paso 15 Para $j = i + 1, i + 2, \ldots, N$ haga los pasos 16-18. *(Efectúe la eliminación.)*

Paso 16 Tome $xm = \dfrac{b_{j,i}}{b_{i,i}}$.

Paso 17 Para $k = i + 1, i + 2, \ldots, N + 1$
tome $b_{j,k} = b_{j,k} - xm \cdot b_{i,k}$.

Paso 18 Tome $b_{j,i} = 0$.

Paso 19 Si $b_{N,N} = 0$ entonces SALIDA ("El sistema es singular");
PARAR.

Paso 20 Si $m > 0$ entonces tome $q_m = \dfrac{b_{N,N+1}}{b_{N,N}}$. *(Comience la sustitución hacia atrás.)*

Paso 21 Para $i = N - 1, N - 2, \ldots, n + 1$ tome $q_{i-n} = \dfrac{b_{i,N+1} - \sum_{j=i+1}^{N} b_{i,j} q_{j-n}}{b_{i,i}}$.

Paso 22 Para $i = n, n-1, \ldots, 1$ tome $p_i = b_{i,N+1} - \sum_{j=n+1}^{N} b_{i,j}\, q_{j-n}$.

Paso 23 SALIDA $(q_0, q_1, \ldots, q_m, p_0, p_1, \ldots, p_n)$;
PARAR. (*Procedimiento terminado exitosamente.*) ■

Es interesante comparar la cantidad de operaciones aritméticas necesarias para calcular $P_5(x)$ y $r(x)$ en el ejemplo 1. Si usamos la multiplicación anidada, $P_5(x)$ puede expresarse como

$$P_5(x) = \left(\left(\left(\left(-\frac{1}{120}x + \frac{1}{24}\right)x - \frac{1}{6}\right)x + \frac{1}{2}\right)x - 1\right)x + 1.$$

Suponiendo que los coeficientes de $1, x, x^2, x^3, x^4$ y x^5 se representan como decimales, un solo cálculo de $P_5(x)$ en forma anidada requiere cinco multiplicaciones y cinco sumas/restas.

Al utilizar la multiplicación anidada, $r(x)$ se expresa así

$$r(x) = \frac{\left(\left(-\frac{1}{60}x + \frac{3}{20}\right)x - \frac{3}{5}\right)x + 1}{\left(\frac{1}{20}x + \frac{2}{5}\right)x + 1},$$

y, por tanto, un solo cálculo de $r(x)$ requiere cinco multiplicaciones, cinco sumas/restas y una división. En consecuencia, el esfuerzo de cálculo requerido parece favorecer a la aproximación polinomial. Pero, al reexpresar $r(x)$ mediante la división continua, podemos escribir

$$\begin{aligned} r(x) &= \frac{1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3}{1 + \frac{2}{5}x + \frac{1}{20}x^2} \\ &= \frac{-\frac{1}{3}x^3 + 3x^2 - 12x + 20}{x^2 + 8x + 20} \\ &= -\frac{1}{3}x + \frac{17}{3} + \frac{\left(-\frac{152}{3}x - \frac{280}{3}\right)}{x^2 + 8x + 20} \\ &= -\frac{1}{3}x + \frac{17}{3} + \frac{-\frac{152}{3}}{\left(\frac{x^2+8x+20}{x+(35/19)}\right)} \end{aligned}$$

o bien

$$r(x) = -\frac{1}{3}x + \frac{17}{3} + \frac{-\frac{152}{3}}{\left(x + \frac{117}{19} + \frac{3125/361}{(x+(35/19))}\right)}. \tag{8.16}$$

Escrito de esta forma, un cálculo simple de $r(x)$ requiere una multiplicación, cinco sumas/restas y dos divisiones. Si la cantidad de cálculos necesarios en la división es aproximadamente igual a la que se requiere en la multiplicación, el esfuerzo de cálculo necesario para evaluar $P_5(x)$ es significativamente mayor que el que se realiza al evaluar $r(x)$.

Se da el nombre de aproximación por **fracción continua**, al hecho de expresar una aproximación de función racional en una forma como la de la ecuación (8.16). Es una técnica clásica de aproximación de gran interés actual, porque esta representación es eficien-

te desde el punto de vista de cálculos requeridos. Sin embargo, se trata de una técnica especializada de la cual no nos ocuparemos más. Un tratamiento bastante exhaustivo de este tema y de la aproximación racional en general se encuentra en [RR, pp. 285-322].

Aunque la aproximación de la función racional del ejemplo 1 dio resultados mejores que los de la aproximación polinomial del mismo grado, la aproximación presenta una exactitud extremadamente variable. La aproximación en 0.2 tiene una exactitud de 8×10^{-9}, mientras que en 1.0 la aproximación y la función concuerdan apenas con una exactitud de 7×10^{-5}. Esta variación de la exactitud es previsible, porque la aproximación de Padé se basa en la representación de e^{-x}, con el polinomio de Taylor, y esta representación presenta una variación de exactitud muy amplia en [0.2, 1.0].

Para obtener una aproximación de la función racional que sea uniformemente más exacta, utilizamos los polinomios de Chebyshev. El método general de la aproximación de Chebyshev a la función racional funciona de la misma manera que la aproximación de Padé, salvo que todo término x^k en la primera se reemplaza con el k-ésimo polinomio $T_k(x)$ de Chebyshev.

Supóngase que queremos aproximar la función f mediante una función racional r de N-ésimo grado, escrita en la forma

$$r(x) = \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}, \quad \text{cuando } N = n + m \text{ y } q_0 = 1.$$

Al escribir $f(x)$ en una serie que contenga los polinomios de Chevyshev, obtenemos

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x)$$

dado

$$f(x) - r(x) = \sum_{k=0}^{\infty} a_k T_k(x) - \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}$$

o bien

$$f(x) - r(x) = \frac{\sum_{k=0}^{\infty} a_k T_k(x) \sum_{k=0}^{m} q_k T_k(x) - \sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}. \tag{8.17}$$

Se eligen los coeficientes $q_1, q_2, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$ de modo que el numerador del lado derecho de esta ecuación tenga coeficientes cero para $T_k(x)$ cuando $k = 0, 1, \ldots, N$. Esto significa que

$$(a_0T_0(x) + a_1T_1(x) + \cdots)(T_0(x) + q_1T_1(x) + \cdots + q_mT_m(x))$$
$$-(p_0T_0(x) + p_1T_1(x) + \cdots + p_nT_n(x))$$

no tiene términos de un grado menor o igual que N.

En el procedimiento de Chebyshev surgen dos problemas que lo hacen un método más difícil de implantar que el método de Padé. Uno de ellos se da porque el producto del polinomio $q(x)$ y la serie de $f(x)$ contienen productos de los polinomios de Chebyshev. Este problema se resuelve utilizando la relación

$$T_i(x)\, T_j(x) = \frac{1}{2}\, [T_{i+j}(x) + T_{|i-j|}(x)]. \tag{8.18}$$

(Véase el ejercicio 8 de la sección 8.3.) El otro problema es más difícil de resolver y requiere el cálculo de la serie de Chebyshev para $f(x)$. En teoría, no es más fácil porque si

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x),$$

entonces la ortogonalidad de los polinomios de Chebyshev implica que

$$a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \quad \text{y} \quad a_k = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)\,T_k(x)}{\sqrt{1-x^2}}\,dx, \quad \text{donde } k \geq 1.$$

No obstante, en la práctica estas integrales rara vez se evalúan en forma cerrada, y para cada evaluación se requiere un método de integración numérica.

EJEMPLO 2 Los cinco primeros términos de la expansión de Chebyshev para e^{-x} son

$$\tilde{P}_5(x) = 1.266066T_0(x) - 1.130318T_1(x) + 0.271495T_2(x) - 0.044337T_3(x) + 0.005474T_4(x) - 0.000543T_5(x).$$

Para determinar la aproximación racional de quinto grado de Chebyshev con $n = 3$ y $m = 2$ se requiere seleccionar p_0, p_1, p_2, p_3, q_1 y q_2 tales que para $k = 0, 1, 2, 3, 4$ y 5 los coeficientes de $T_k(x)$ sean cero en la expansión

$$\tilde{P}_5(x)[T_0(x) + q_1T_1(x) + q_2T_2(x)] - [p_0T_0(x) + p_1T_1(x) + p_2T_2(x) + p_3T_3(x)].$$

Al usar la relación (8.18) y al reunir los términos, obtenemos las ecuaciones

$$\begin{aligned}
T_0:&\quad 1.266066 - 0.565159q_1 + 0.1357485q_2 = p_0,\\
T_1:&\quad -1.130318 + 1.401814q_1 - 0.587328q_2 = p_1,\\
T_2:&\quad 0.271495 - 0.587328q_1 + 1.268803q_2 = p_2,\\
T_3:&\quad -0.044337 + 0.138485q_1 - 0.565431q_2 = p_3,\\
T_4:&\quad 0.005474 - 0.022440q_1 + 0.135748q_2 = 0,\\
T_5:&\quad -0.000543 + 0.002737q_1 - 0.022169q_2 = 0.
\end{aligned}$$

La solución de este sistema genera la función racional

$$r_T(x) = \frac{1.055265T_0(x) - 0.613016T_1(x) + 0.077478T_2(x) - 0.004506T_3(x)}{T_0(x) + 0.378331T_1(x) + 0.022216T_2(x)}$$

Al inicio de la sección 8.3 encontramos que

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x.$$

Al usar esos valores para convertir la expresión anterior en una expresión que contenga potencias de x, obtenemos

$$r_T(x) = \frac{0.977787 - 0.599499x + 0.154956x^2 - 0.018022x^3}{0.977784 + 0.378331x + 0.044432x^2}.$$

La tabla 8.11 contiene los valores de $r_T(x)$ y para facilitar la comparación los valores de $r(x)$ que se obtuvieron en el ejemplo 1. Obsérvese que la aproximación dada por $r(x)$ es superior a la de $r_T(x)$ para $x = 0.2$ y 0.4; pero el error máximo de $r(x)$ es 6.33×10^{-5} en comparación con 9.13×10^{-6} para $r_T(x)$. ■

Tabla 8.11

x	e^{-x}	$r(x)$	$\lvert e^{-x} - r(x)\rvert$	$r_T(x)$	$\lvert e^{-x} - r_T(x)\rvert$
0.2	0.81873075	0.81873075	7.55×10^{-9}	0.81872510	5.66×10^{-6}
0.4	0.67032005	0.67031963	4.11×10^{-7}	0.67031310	6.95×10^{-6}
0.6	0.54881164	0.54880763	4.00×10^{-6}	0.54881292	1.28×10^{-6}
0.8	0.44932896	0.44930966	1.93×10^{-5}	0.44933809	9.13×10^{-6}
1.0	0.36787944	0.36781609	6.33×10^{-5}	0.36787155	7.89×10^{-6}

La aproximación de Chebyshev puede generarse por medio del algoritmo 8.2.

ALGORITMO 8.2

Aproximación racional de Chebyshev

Para obtener la aproximación racional

$$r_T(x) = \frac{\sum_{k=0}^{n} p_k T_k(x)}{\sum_{k=0}^{m} q_k T_k(x)}$$

de una función determinada $f(x)$:

ENTRADA enteros no negativos m y n.

SALIDA coeficientes $q_0, q_1, \ldots, q_m$ y $p_0, p_1, \ldots, p_n$.

Paso 1 Tome $N = m + n$.

Paso 2 Tome $a_0 = \frac{2}{\pi} \int_0^{\pi} f(\cos \theta)\, d\theta$; *(El coeficiente a_0 es doble para lograr la eficiencia en el proceso de cálculo.)*

Para $k = 1, 2, \ldots, N + m$ tome

$$a_k = \frac{2}{\pi} \int_0^{\pi} f(\cos \theta) \cos k\theta\, d\theta.$$

(Las integrales se pueden evaluar usando un procedimiento de integración numérica o los coeficientes pueden introducirse directamente.)

Paso 3 Tome $q_0 = 1$.

Paso 4 Para $i = 0, 1, \ldots, N$ haga pasos 5-9. *(Utilice un sistema lineal con matriz B.)*

Paso 5 Para $j = 0, 1, \ldots, i$
si $j \leq n$ entonces tome $b_{i,j} = 0$.

Paso 6 Si $i \leq n$ entonces tome $b_{i,i} = 1$.

Paso 7 Para $j = i + 1, i + 2, \ldots, n$ tome $b_{i,j} = 0$.

Paso 8 Para $j = n + 1, n + 2, \ldots, N$

si $i \neq 0$ entonces tome $b_{i,j} = -\frac{1}{2}(a_{i+j-n} + a_{|i-j+n|})$
si no, tome $b_{i,j} = -\frac{1}{2}a_{j-n}$.

Paso 9 Si $i \neq 0$ entonces tome $b_{i,N+1} = a_i$
si no, tome $b_{i,N+1} = \frac{1}{2}a_i$.

(*Los pasos 10-21 resuelven el sistema lineal usando el pivoteo parcial.*)

Paso 10 Para $i = n + 1, n + 2, \ldots, N - 1$ haga los pasos 11-17.

Paso 11 Sea k el entero más pequeño con $i \leq k \leq N$ y
$|b_{k,i}| = \text{máx}_{i\leq j\leq N} |b_{j,i}|$. (*Obtenga el elemento de pivote.*)

Paso 12 Si $b_{k,i} = 0$ entonces SALIDA ("El sistema es singular".)
PARAR.

Paso 13 Si $k \neq i$ entonces (*Intercambie el renglón i y el renglón k.*)
para $j = i, i + 1, \ldots N + 1$ tome
$b_{COPY} = b_{i,j}$;
$b_{i,j} = b_{k,j}$;
$b_{k,j} = b_{COPY}$.

Paso 14 Para $j = i + 1, i + 2, \ldots, N$ haga los pasos 15-17. (*Efectúe la eliminación.*)

Paso 15 Tome $xm = \dfrac{b_{j,i}}{b_{i,i}}$.

Paso 16 Para $k = i + 1, i + 2, \ldots, N + 1$
tome $b_{j,k} = b_{j,k} - xm \cdot b_{i,k}$.

Paso 17 Tome $b_{j,i} = 0$.

Paso 18 Si $b_{N,N} = 0$ entonces SALIDA ("El sistema es singular");
PARAR.

Paso 19 Si $m > 0$ entonces tome $q_m = \dfrac{b_{N,N+1}}{b_{N,N}}$. (*Comience la sustitución hacia atrás.*)

Paso 20 Para $i = N-1, N-2, \ldots, n + 1$ tome $q_{i-n} = \dfrac{b_{i,N+1} - \sum_{j=i+1}^{N} b_{i,j}q_{j-n}}{b_{i,i}}$.

Paso 21 Para $i = n, n-1, \ldots, 0$ tome $p_i = b_{i,N+1} - \sum_{j=n+1}^{N} b_{i,j}q_{j-n}$.

Paso 22 SALIDA $(q_0, q_1, \ldots, q_m, p_0, p_1, \ldots, p_n)$;
PARAR. (*Procedimiento terminado exitosamente.*) ■

Mediante el uso de Maple podemos obtener la expansión de la serie de Chebyshev y la aproximación racional de Chebyshev. Para hacer los polinomios de Chebyshev accesibles a Maple se introduce el comando

```
>with(orthopoly,T);
```

El procedimiento para calcular la serie de Chebyshev como una aproximación es

```
>g:=numapprox[chebyshev](exp(-x),x,0.000001);
```

donde el tercer parámetro especifica la exactitud requerida. El resultado es

$$
\begin{aligned}
g := {} & 1.266065878\, T(0, x) - 1.130318208\, T(1, x) + .2714953396\, T(2, x) \\
& -.04433684985\, T(3, x) + .005474240442\, T(4, x) \\
& - .0005429263119\, T(5, x) + .00004497732296\, T(6, x) \\
& - .3198436462\, 10^{-5}\, T(7, x)
\end{aligned}
$$

y podemos evaluar $g(0.8)$ utilizando

```
>evalf(subs(x=0.8,g));
```

para obtener .4493288893.

Para obtener la aproximación racional de Chebyshev, partimos de la serie de Chebyshev

```
>restart;
>numapprox[chebyshev](exp(-x),x,0.000001);
```

como antes, y luego introducimos

```
>g:=convert(%,ratpoly,3,2);
```

que da por resultado

$$
\begin{aligned}
g := {} & (1.050531166\, T(0, x) - .6016362122\, T(1, x) + .07417897149\, T(2, x) \\
& -.004109558353\, T(3, x))/(T(0, x) + .3870509565\, T(1, x) \\
& + .02365167312\, T(2, x))
\end{aligned}
$$

Como ya vaciamos la memoria de Maple, necesitamos reintroducir el comando

```
>with(orthopoly,T);
```

para que podamos evaluar $g(0.8)$ mediante

```
>evalf(subs(x=0.8,g));
```

y obtener .4493317579.

El método de Chebyshev no produce la mejor aproximación de la función racional, en el sentido de la aproximación cuyo error máximo sea el menor posible. No obstante, podemos usar este método como punto de arranque de un procedimiento iterativo denominado segundo algoritmo de Remes que converge a la mejor aproximación. En [RR, pp. 292-305] y en [Po, pp. 90-92] encontramos una explicación de las técnicas de que consta este método y un perfeccionamiento de este algoritmo.

CONJUNTO DE EJERCICIOS 8.4

1. Determine todas las aproximaciones de Padé de segundo grado para $f(x) = e^{2x}$. Compare los resultados en $x_i = 0.2i$, para $i = 1, 2, 3, 4, 5$, con los valores reales $f(x_i)$.
2. Determine todas las aproximaciones de Padé de tercer grado para $f(x) = x\ln(x+1)$. Compare los resultados en $x_i = 0.2i$, para $i = 1, 2, 3, 4, 5$, con los valores reales $f(x_i)$.
3. Determine la aproximación de Padé de quinto grado con $n = 2$ y $m = 3$ para $f(x) = e^x$. Compare los resultados en $x_i = 0.2i$, para $i = 1, 2, 3, 4, 5$, con los del quinto polinomio de Maclaurin.
4. Repita el ejercicio 3 usando esta vez la aproximación de Padé de quinto grado con $n = 3$ y $m = 2$. Compare los resultados en x_i con los que se obtuvieron en el ejercicio 3.
5. Determine la aproximación de Padé de sexto grado con $n = m = 3$ para $f(x) = \operatorname{sen} x$. Compare los resultados en $x_i = 0.1i$, para $i = 0, 1, \ldots, 5$, con los resultados exactos y con los del sexto polinomio de Maclaurin.
6. Determine las aproximaciones de sexto grado de Padé con (a) $n = 2$, $m = 4$ y (b) $n = 4$, $m = 2$ para $f(x) = \operatorname{sen} x$. Compare los resultados en cada x_i con los obtenidos en el ejercicio 5.
7. La tabla 8.10 contiene los resultados de la aproximación de Padé de quinto grado con $n = 3$ y $m = 2$, el quinto polinomio de Maclaurin, y los valores exactos de $f(x) = e^{-x}$ cuando $x_i = 0.2i$, para $i = 1, 2, 3, 4$ y 5. Compare estos resultados con los producidos en otras aproximaciones de Padé de quinto grado.

   a. $n = 0, m = 5$
   b. $n = 1, m = 4$
   c. $n = 3, m = 2$
   d. $n = 4, m = 1$

8. Exprese las siguientes funciones racionales en forma de fracción continua:

   a. $\dfrac{x^2 + 3x + 2}{x^2 - x + 1}$
   b. $\dfrac{4x^2 + 3x - 7}{2x^3 + x^2 - x + 5}$
   c. $\dfrac{2x^3 - 3x^2 + 4x - 5}{x^2 + 2x + 4}$
   d. $\dfrac{2x^3 + x^2 - x + 3}{3x^3 + 2x^2 - x + 1}$

9. Obtenga todas las aproximaciones racionales de Chebyshev de segundo grado para $f(x) = e^{-x}$. ¿Cuáles ofrecen las mejores aproximaciones a $f(x) = e^{-x}$ en $x = 0.25, 0.5$ y 1?
10. Obtenga todas las aproximaciones racionales de Chebyshev de tercer grado para $f(x) = \cos x$. ¿Cuáles ofrecen las mejores aproximaciones a $f(x) = \cos x$ para $x = \pi/4$ y $\pi/3$?
11. Obtenga la aproximación racional de Chebyshev de cuarto grado con $n = m = 2$ para $f(x) = \operatorname{sen} x$. Compare los resultados en $x_i = 0.1i$, para $i = 0, 1, 2, 3, 4, 5$ con los obtenidos en el ejercicio 5 mediante la aproximación de Padé de sexto grado.
12. Obtenga todas las aproximaciones racionales de Chebyshev de quinto grado para $f(x) = e^x$. Compare los resultados en $x_i = 0.2i$, para $i = 1, 2, 3, 4, 5$ obtenidos con esta aproximación con los obtenidos en los ejercicios 3 y 4.
13. Si queremos aproximar exactamente $f(x) = e^x$ para su inclusión en una biblioteca de matemáticas, primero restringimos el dominio de f. Dado un número real x, divídalo entre $\ln\sqrt{10}$ para obtener la relación

$$x = M \cdot \ln\sqrt{10} + s,$$

donde M es un entero y s es un número real que satisface $|s| \le \frac{1}{2}\ln\sqrt{10}$.

   a. Demuestre que $e^x = e^s \cdot 10^{M/2}$.

b. Construya una aproximación de la función racional para e^s usando $n = m = 3$. Estime el error cuando $0 \leq |s| \leq \frac{1}{2} \ln \sqrt{10}$.

c. Diseñe una implantación de e^x usando los resultados de las partes (a) y (b) y las aproximaciones

$$\frac{1}{\ln \sqrt{10}} = 0.8685889638 \quad \text{y} \quad \sqrt{10} = 3.162277660.$$

14. Si queremos aproximar exactamente sen x y cos x para su inclusión en una biblioteca de matemáticas, primero restringimos sus dominios. Dado un número real x, divídalo entre π para obtener la relación

$$|x| = M\pi + s, \quad \text{donde } M \text{ es una integral y } |s| \leq \frac{\pi}{2}.$$

a. Demuestre que sen $x = \text{sgn}(x) \cdot (-1)^M \cdot$ sen s.

b. Construya una aproximación racional a sen s usando $n = m = 4$. Estime el error cuando $0 \leq |s| \leq \pi/2$.

c. Diseñe una implantación de sen x usando las partes (a) y (b).

d. Repita la parte (c) para cos x aplicando el hecho de que cos x = sen$(x + \pi/2)$.

8.5 Aproximación polinomial trigonométrica

El uso de las series de las funciones seno y coseno para representar funciones arbitrarias comenzó en la década de 1750, con el estudio del movimiento de un resorte en vibración. Este problema fue estudiado por Jean d'Alembert y luego también por el matemático más destacado de la época, Leonhard Euler. Pero fue Daniel Bernoulli el primero en proponer el uso de sumas infinitas de senos y cosenos como solución al problema, sumas que ahora conocemos con el nombre de series de Fourier. A principios del siglo XIX, Jean Baptiste Joseph Fourier las utilizó para estudiar el flujo de calor y formuló una teoría muy completa sobre el tema.

La primera observación en el desarrollo de las series de Fourier es que, para todo entero positivo n, el conjunto de las funciones $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ donde

$$\phi_0(x) = \frac{1}{2},$$

$$\phi_k(x) = \cos kx, \quad \text{para cada } k = 1, 2, \ldots, n,$$

y

$$\phi_{n+k}(x) = \text{sen } kx, \quad \text{para cada } k = 1, 2, \ldots, n-1,$$

es un conjunto ortogonal en $[-\pi, \pi]$ con respecto a $w(x) \equiv 1$. Esta ortogonalidad se deduce del hecho de que, para todo entero j, las integrales de sen jx y cos jx en $[-\pi, \pi]$ son 0, y podemos reescribir los productos de las funciones seno y coseno como sumas utilizando las tres identidades trigonométricas

$$\text{sen } t_1 \text{ sen } t_2 = \frac{1}{2}[\cos(t_1 - t_2) - \cos(t_1 + t_2)],$$

$$\cos t_1 \cos t_2 = \frac{1}{2}[\cos(t_1 - t_2) + \cos(t_1 + t_2)], \qquad (8.19)$$

$$\text{sen } t_1 \cos t_2 = \frac{1}{2}[\text{sen}(t_1 - t_2) + \text{sen}(t_1 + t_2)].$$

Sea $\mathcal{T}_n$ el conjunto de todas las combinaciones lineales de las funciones $\phi_0, \phi_1, \ldots, \phi_{2n-1}$. A este conjunto se le denomina conjunto de **polinomios trigonométricos** de grado menor o igual que n. (En algunas obras se incluye una función adicional en el conjunto, $\phi_{2n}(x) = \text{sen } nx$.)

Para una función $f \in C[-\pi, \pi]$, queremos obtener la aproximación de *mínimos cuadrados continuos* mediante las funciones de $\mathcal{T}_n$ en la forma

$$S_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1} (a_k \cos kx + b_k \text{ sen } kx).$$

Dado que el conjunto de funciones $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ es ortogonal en $[-\pi, \pi]$ respecto a $w(x) = 1$, del teorema 8.6 se deduce que la elección apropiada de coeficientes es

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos kx \, dx, \qquad \text{para cada } k = 0, 1, 2, \ldots, n,$$

y

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \text{ sen } kx \, dx, \qquad \text{para cada } k = 1, 2, \ldots, n-1.$$

El límite de $S_n(x)$ cuando $n \to \infty$ se denomina **serie de Fourier** para f. Las series de Fourier se usan para describir la solución de varias ecuaciones diferenciales, ordinarias y parciales, que aparecen en situaciones físicas.

EJEMPLO 1 Para determinar el polinomio trigonométrico a partir de $\mathcal{T}_n$ que aproxime a

$$f(x) = |x|, \qquad \text{para } -\pi < x < \pi,$$

hay que encontrar

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \, dx = -\frac{1}{\pi}\int_{-\pi}^{0} x \, dx + \frac{1}{\pi}\int_{0}^{\pi} x \, dx = \frac{2}{\pi}\int_{0}^{\pi} x \, dx = \pi,$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \cos kx \, dx = \frac{2}{\pi}\int_{0}^{\pi} x \cos kx \, dx = \frac{2}{\pi k^2}\,[(-1)^k - 1],$$

para cada $k = 1, 2, \ldots, n$; y

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \text{ sen } kx \, dx = 0, \qquad \text{para cada } k = 1, 2, \ldots, n-1.$$

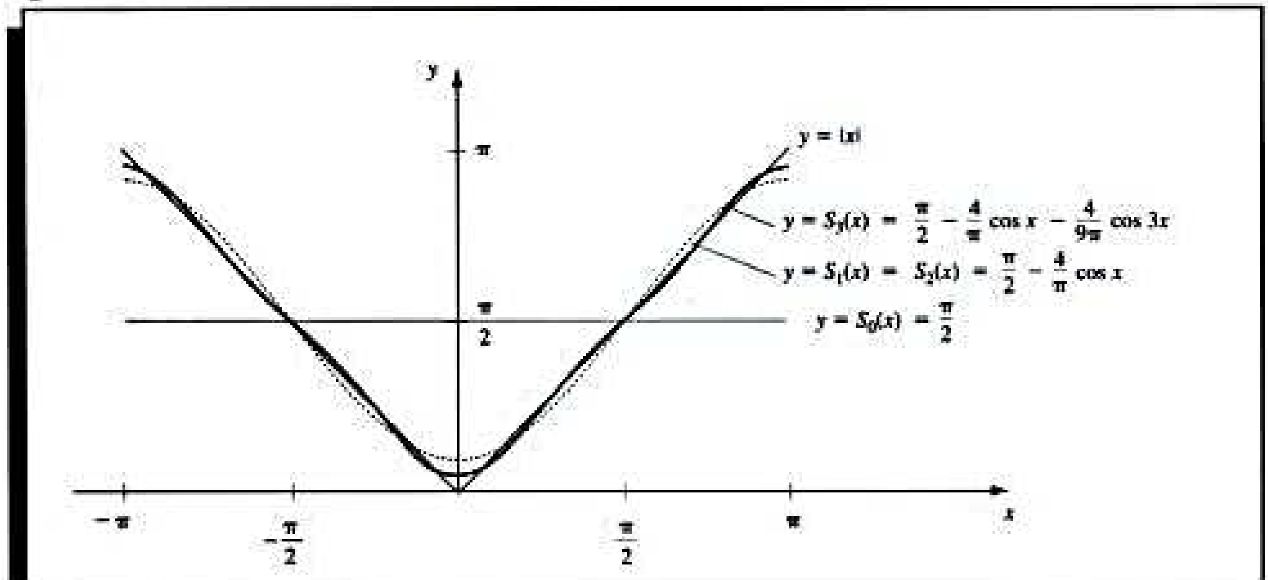
El que las b_k sean todas 0 se deduce del hecho de que $g(x) = |x| \text{ sen } kx$ es una función impar para cada k, y la integral de cualquier función impar en todo intervalo de la forma $[-a, a]$ es 0. (Véanse los Ejercicios 13 y 14.) Así pues, el polinomio trigonométrico obte-

nido de $\mathcal{T}_n$ que aproxima f es

$$S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{n} \frac{(-1)^k - 1}{k^2} \cos kx.$$

En la figura 8.13 se muestran algunos de los primeros polinomios trigonométricos para $f(x) = |x|$.

Figura 8.13



La serie de Fourier para f es

$$S(x) = \lim_{n\to\infty} S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{\infty} \frac{(-1)^k - 1}{k^2} \cos kx.$$

Como $|\cos kx| \leq 1$, para toda k y x, la serie converge, y $S(x)$ existe para todo número real x. ∎

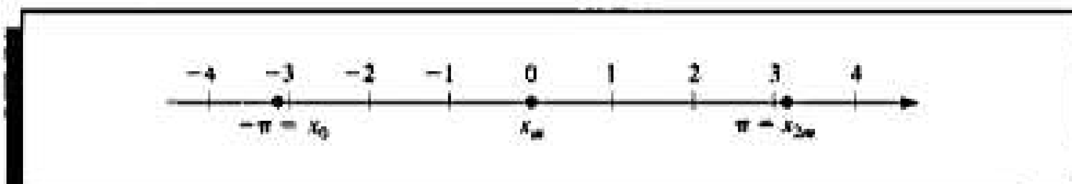
Hay un análogo discreto de gran utilidad para la aproximación de *mínimos cuadrados discretos* y para la interpolación de grandes cantidades de datos.

Supóngase que tenemos un conjunto de $2m$ puntos de parejas de datos $\{(x_j, y_j)\}_{j=0}^{2m-1}$ y que los primeros elementos de los pares dividen uniformemente un intervalo cerrado. Para simplificar la explicación supondremos que el intervalo es $[-\pi, \pi]$ así que, como se observa en la figura 8.14,

$$x_j = -\pi + \left(\frac{j}{m}\right)\pi, \quad \text{para cada } j = 0, 1, \ldots, 2m - 1. \tag{8.20}$$

Si el intervalo no es $[-\pi, \pi]$, podríamos emplear una transformación lineal simple para traducir los datos en esta forma.

Figura 8.14



El objetivo del caso discreto consiste en determinar el polinomio trigonométrico $S_n(x)$ en $\mathcal{T}_n$ que reduzca al mínimo

$$E(S_n) = \sum_{j=0}^{2m-1} [y_j - S_n(x_j)]^2.$$

Para ello necesitamos seleccionar las constantes $a_0, a_1, \ldots, a_n, b_1, b_2, \ldots, b_{n-1}$ de modo que

$$E(S_n) = \sum_{j=0}^{2m-1} \left\{ y_j - \left[\frac{a_0}{2} + a_n \cos nx_j + \sum_{k=1}^{n-1} (a_k \cos kx_j + b_k \operatorname{sen} kx_j) \right] \right\}^2 \tag{8.21}$$

sea un mínimo.

La determinación de las constantes se simplifica por el hecho de que el conjunto $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ también es ortogonal respecto a la suma en puntos uniformemente espaciados $\{x_j\}_{j=0}^{2m-1}$ en $[-\pi, \pi]$. Con ello queremos decir que para cada $k \neq l$,

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_l(x_j) = 0. \tag{8.22}$$

Para demostrar esta ortogonalidad recurrimos al siguiente lema.

Lema 8.12 Si el entero r no es un múltiplo de $2m$, entonces

$$\sum_{j=0}^{2m-1} \cos rx_j = 0 \quad \text{y} \quad \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = 0.$$

Más aún, si r no es un múltiplo de m, entonces

$$\sum_{j=0}^{2m-1} (\cos rx_j)^2 = m \quad \text{y} \quad \sum_{j=0}^{2m-1} (\operatorname{sen} rx_j)^2 = m. \qquad \blacksquare$$

Demostración La *fórmula de Euler* establece que si $i^2 = -1$, entonces para cada número real z, tenemos

$$e^{iz} = \cos z + i \operatorname{sen} z. \tag{8.23}$$

Al aplicar este resultado nos da

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = \sum_{j=0}^{2m-1} (\cos rx_j + i \operatorname{sen} rx_j) = \sum_{j=0}^{2m-1} e^{irx_j}.$$

Pero

$$e^{irx_j} = e^{ir(-\pi + j\pi/m)} = e^{-ir\pi} \cdot e^{irj\pi/m},$$

así que

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = e^{-ir\pi} \sum_{j=0}^{2m-1} e^{irj\pi/m}.$$

Como $\sum_{j=0}^{2m-1} e^{irj\pi/m}$ es una serie geométrica con el primer término 1 y con la razón $e^{ir\pi/m} \neq 1$, tenemos

$$\sum_{j=0}^{2m-1} e^{irj\pi/m} = \frac{1 - (e^{ir\pi/m})^{2m}}{1 - e^{ir\pi/m}} = \frac{1 - e^{2ir\pi}}{1 - e^{ir\pi/m}}.$$

Pero $e^{2ir\pi} = \cos 2r\pi + i \operatorname{sen} 2r\pi = 1$, así que $1 - e^{2ir\pi} = 0$ y

$$\sum_{j=0}^{2m-1} \cos rx_j + i \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = e^{-ir\pi} \sum_{j=0}^{2m-1} e^{irj\pi/m} = 0.$$

Esto implica que

$$\sum_{j=0}^{2m-1} \cos rx_j = 0 \quad \text{y} \quad \sum_{j=0}^{2m-1} \operatorname{sen} rx_j = 0.$$

Si r no es un múltiplo de m, estas sumas implican que

$$\sum_{j=0}^{2m-1} (\cos rx_j)^2 = \sum_{j=0}^{2m-1} (1 + \cos 2rx_j)$$
$$= \frac{1}{2}\left[\sum_{j=0}^{2m-1} 1 + \sum_{j=0}^{2m-1} \cos 2rx_j\right] = \frac{1}{2}(2m + 0) = m$$

y, de manera semejante, que

$$\sum_{j=0}^{2m-1} (\operatorname{sen} rx_j)^2 = m. \qquad \blacksquare\ \blacksquare\ \blacksquare$$

Ahora podemos demostrar la ortogonalidad establecida en (8.22). Consideremos, por ejemplo, el caso

$$\sum_{j=0}^{2m-1} \phi_k(x_j)\phi_{n+l}(x_j) = \sum_{j=0}^{2m-1} (\cos kx_j)(\operatorname{sen} lx_j).$$

Dado que

$$\cos kx_j \operatorname{sen} lx_j = \frac{1}{2}\left[\operatorname{sen}(l + k)x_j + \operatorname{sen}(l - k)x_j\right]$$

y que $(l + k)$ y $(l - k)$ son enteros no múltiplos de $2m$, el lema 8.12 implica que

$$\sum_{j=0}^{2m-1} (\cos kx_j)(\operatorname{sen} lx_j) = \frac{1}{2}\left[\sum_{j=0}^{2m-1} \operatorname{sen}(l+k)x_j + \sum_{j=0}^{2m-1} \operatorname{sen}(l-k)x_j\right] = \frac{1}{2}(0+0) = 0.$$

Este método se usa para demostrar que la condición de ortogonalidad se satisface para cualquier par de las funciones y se emplea para producir el siguiente resultado.

Teorema 8.13 Las constantes de la sumatoria

$$s_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1} (a_k \cos kx + b_k \operatorname{sen} kx)$$

que reducen al mínimo la suma de mínimos cuadrados

$$E(a_0, \ldots, a_n, b_1, \ldots, b_{n-1}) = \sum_{j=0}^{2m-1} (y_j - S_n(x_j))^2$$

son

$$a_k = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \cos kx_j, \quad \text{para cada } k = 0, 1, \ldots, n,$$

y

$$b_k = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j, \quad \text{para cada } k = 1, 2, \ldots, n-1.$$ ■

Este teorema se demuestra igualando a cero las derivadas parciales de E respecto a las a_k y las b_k, como se hizo en las secciones 8.1 y 8.2, y aplicando la ortogonalidad para simplificar las ecuaciones. Por ejemplo,

$$0 = \frac{\partial E}{\partial b_k} = 2 \sum_{j=0}^{2m-1} [y_j - S_n(x_j)](-\operatorname{sen} kx_j),$$

así que

$$\begin{aligned} 0 &= \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j - \sum_{j=0}^{2m-1} S_n(x_j) \operatorname{sen} kx_j \\ &= \sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j - \frac{a_0}{2} \sum_{j=0}^{2m-1} \operatorname{sen} kx_j - a_n \sum_{j=0}^{2m-1} \operatorname{sen} kx_j \cos nx_j \\ &\quad - \sum_{l=1}^{n-1} a_l \sum_{j=0}^{2m-1} \operatorname{sen} kx_j \cos lx_j - \sum_{\substack{l=1,\\ l\neq k}}^{n-1} b_l \sum_{j=0}^{2m-1} \operatorname{sen} kx_j \operatorname{sen} lx_j - b_k \sum_{j=0}^{2m-1} (\operatorname{sen} kx_j)^2. \end{aligned}$$

La ortogonalidad implica que todas las sumas del lado derecho menos la primera y última son cero, y el lema 8.12 establece que la suma final es m. Por tanto,

$$b_k = \frac{1}{m}\sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j.$$

EJEMPLO 2 Sea $f(x) = x^4 - 3x^3 + 2x^2 - \tan x(x-2)$. Para obtener la aproximación de mínimos cuadrados discretos $S_3(x)$ para los datos $\{(x_j, y_j)\}_{j=0}^{9}$, donde $x_j = j/5$ y $y_j = f(x_j)$, se requiere una transformación de $[0, 2]$ a $[-\pi, \pi]$. Es fácil verificar que la transformación lineal requerida de la siguiente manera:

$$z_j = \pi(x_j - 1),$$

que los datos traducidos presentan la forma

$$\left\{\left(z_j, f\left(1 + \frac{z_j}{\pi}\right)\right)\right\}_{j=0}^{9}.$$

En consecuencia, el polinomio trigonométrico de mínimos cuadrados es

$$S_3(z) = \left[\frac{a_0}{2} + a_3 \cos 3z + \sum_{k=1}^{2}(a_k \cos kz + b_k \operatorname{sen} kz)\right],$$

donde

$$a_k = \frac{1}{5}\sum_{j=0}^{9} f\left(1 + \frac{z_j}{\pi}\right)\cos kz_j, \quad \text{para } k = 0, 1, 2, 3,$$

y

$$b_k = \frac{1}{5}\sum_{j=0}^{9} f\left(1 + \frac{z_j}{\pi}\right)\operatorname{sen} kz_j, \quad \text{para } k = 1, 2.$$

Al evaluar estas sumas obtenemos la aproximación

$$S_3(z) = 0.76201 + 0.77177\cos z + 0.017423\cos 2z + 0.0065673\cos 3z$$
$$-0.38676 \operatorname{sen} z + 0.047806 \operatorname{sen} 2z,$$

Tabla 8.12

x	$f(x)$	$S_3(x)$	$\lvert f(x) - S_3(x)\rvert$
0.125	0.26440	0.24060	2.38×10^{-2}
0.375	0.84081	0.85154	1.07×10^{-2}
0.625	1.36150	1.36248	9.74×10^{-4}
0.875	1.61282	1.60406	8.75×10^{-3}
1.125	1.36672	1.37566	8.94×10^{-3}
1.375	0.71697	0.71545	1.52×10^{-3}
1.625	0.07909	0.06929	9.80×10^{-3}
1.875	−0.14576	−0.12302	2.27×10^{-2}

Al convertir de nuevo a la variable x, obtenemos

$$s_3(x) = 0.76201 + 0.77177 \cos \pi(x-1) + 0.017423 \cos 2\pi(x-1)$$
$$+ 0.0065673 \cos 3\pi(x-1) - 0.38676 \operatorname{sen} \pi(x-1) + 0.047806 \operatorname{sen} 2\pi(x-1).$$

La tabla 8.12 contiene los valores de $f(x)$ y de $S_3(x)$. ■

CONJUNTO DE EJERCICIOS 8.5

1. Obtenga el polinomio trigonométrico de mínimos cuadrados continuos $S_2(x)$ para $f(x) = x^2$ en $[-\pi, \pi]$.
2. Obtenga el polinomio trigonométrico de mínimos cuadrados continuos $S_n(x)$ para $f(x) = x$ en $[-\pi, \pi]$.
3. Obtenga el polinomio trigonométrico de mínimos cuadrados continuos $S_3(x)$ para $f(x) = e^x$ en $[-\pi, \pi]$.
4. Obtenga el polinomio trigonométrico de mínimos cuadrados continuos $S_n(x)$ para $f(x) = e^x$ en $[-\pi, \pi]$.
5. Obtenga el polinomio trigonométrico general de mínimos cuadrados continuos $S_n(x)$ para

$$f(x) = \begin{cases} 0, & \text{si } -\pi < x \leq 0, \\ 1, & \text{si } 0 < x < \pi. \end{cases}$$

6. Obtenga el polinomio trigonométrico general de mínimos cuadrados continuos $S_n(x)$ para

$$f(x) = \begin{cases} -1, & \text{si } -\pi < x < 0, \\ 1, & \text{si } 0 \leq x \leq \pi. \end{cases}$$

7. Determine el polinomio trigonométrico de mínimos cuadrados discretos $S_n(x)$ en el intervalo $[-\pi, \pi]$ para las siguientes funciones empleando los valores dados de m y n:
   a. $f(x) = \cos 2x, \quad m = 4, n = 2$
   b. $f(x) = \cos 3x, \quad m = 4, n = 2$
   c. $f(x) = \operatorname{sen} \frac{1}{2}x + 2 \cos \frac{1}{3}x, \quad m = 6, n = 3$
   d. $f(x) = x^2 \cos x, \quad m = 6, n = 3.$
8. Calcule el error $E(S_n)$ para las funciones del ejercicio 7.
9. Determine el polinomio trigonométrico de mínimos cuadrados discretos $S_3(x)$, utilizando $m = 4$ para $f(x) = e^x \cos 2x$ en el intervalo $[-\pi, \pi]$. Calcule el error $E(S_3)$.
10. Repita el ejercicio 9 empleando $m = 8$. Compare los valores de los polinomios aproximantes con los de f en los puntos $\xi_j = \pi + 0.2j\pi$, para $0 \leq j \leq 10$. ¿Cuál aproximación es mejor?
11. Sea $f(x) = 2 \tan x - \sec 2x$, para $2 \leq x \leq 4$. Determine los polinomios trigonométricos de mínimos cuadrados discretos $S_n(x)$, empleando los valores de n y m como sigue y después calcule el error en cada caso.
    a. $n = 3, \quad m = 6$
    b. $n = 4, \quad m = 6$
12. a. Determine el polinomio trigonométrico de mínimos cuadrados discretos $S_4(x)$ usando $m = 16$, para $f(x) = x^2 \operatorname{sen} x$ en el intervalo $[0, 1]$.
    b. Calcule $\int_0^1 S_4(x)\,dx$.
    c. Compare la integral de la parte (b) con $\int_0^1 x^2 \operatorname{sen} x\,dx$.

13. Demuestre que para cualquier función impar continua f definida en el intervalo $[-a, a]$, tenemos $\int_{-a}^{a} f(x)\,dx = 0$.
14. Demuestre que para toda función par continua f definida en el intervalo $[-a, a]$, tenemos $\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx$.
15. Demuestre que las funciones $\phi_0(x) = 1/2$, $\phi_1(x) = \cos x, \ldots, \phi_n(x) = \cos nx$, $\phi_{n+1}(x) = \operatorname{sen} x, \ldots, \phi_{2n-1}(x) = \operatorname{sen}(n-1)x$ son ortogonales en $[-\pi, \pi]$ respecto a $w(x) \equiv 1$.
16. En el ejemplo 1, se determinó la serie de Fourier para $f(x) = |x|$. Use esta serie y la suposición de que representa a f en cero, para calcular el valor de la serie infinita convergente $\sum_{k=0}^{\infty} (1/(2k+1)^2)$.

8.6 Transformadas rápidas de Fourier

En la segunda mitad de la sección 8.5 determinamos la forma del polinomio de mínimos cuadrados discretos de n-ésimo grado en los $2m - 1$ puntos de datos $\{(x_j, y_j)\}_{j=0}^{2m-1}$ donde $x_j = -\pi + (j/m)\,\pi$, para cada $j = 0, 1, \ldots, 2m-1$.

El polinomio trigonométrico *interpolante* de S_m en estos $2m$ puntos de datos es casi el mismo que el del polinomio de mínimos cuadrados. Esto se debe a que el polinomio trigonométrico de mínimos cuadrados reduce al mínimo el término de error

$$E(S_m) = \sum_{j=0}^{2m-1} (y_j - S_m(x_j))^2,$$

y este error es cero; de ahí que se disminuya al mínimo cuando $S_m(x_j) = y_j$ para cada $j = 0, 1, \ldots, 2m-1$. Sin embargo, se necesita modificar la forma del polinomio, si queremos que los coeficientes asuman la misma forma que en el caso de los mínimos cuadrados. En el lema 8.12 comprobamos que, si r no es un múltiplo de m, entonces

$$\sum_{j=0}^{2m-1} (\cos rx_j)^2 = m.$$

La interpolación requiere calcular

$$\sum_{j=0}^{2m-1} (\cos mx_j)^2,$$

lo cual (véase el ejercicio 8) tiene el valor $2m$. Para esto es necesario escribir el polinomio interpolante así

$$S_m(x) = \frac{a_0 + a_m \cos mx}{2} + \sum_{j=0}^{m-1} (a_k \cos kx + b_k \operatorname{sen} kx), \tag{8.24}$$

si queremos que la forma de las constantes a_k y b_k concuerde con las del polinomio de mínimos cuadrados discretos. Por tanto, queremos que las constantes sean

$$a_k = \frac{1}{m} \sum_{j=0}^{2m-1} y_j \cos kx_j, \qquad \text{para cada } k = 0, 1, \ldots, m, \tag{8.25}$$

y

$$b_k = \frac{1}{m}\sum_{j=0}^{2m-1} y_j \operatorname{sen} kx_j, \qquad \text{para cada } k = 1, 2, \ldots, m-1. \tag{8.26}$$

La interpolación de grandes cantidades de datos uniformemente espaciados por medio de polinomios trigonométricos, puede dar resultados muy exactos. Es el método adecuado de aproximación que se emplea en áreas como las de filtros digitales, patrones de campo de antena, la mecánica cuántica y la óptica, así como muchas áreas relacionadas con los problemas de simulación. Sin embargo, hasta mediados de los años sesenta el método no tenía gran aplicación debido a que se requerían muchos cálculos aritméticos en la determinación de las constantes de la aproximación. La interpolación de $2m$ puntos de datos mediante la técnica de cálculo directo requiere cerca de $(2m)^2$ multiplicaciones y $(2m)^2$ sumas. La aproximación de miles de datos no es poco común en áreas donde hay que utilizar la interpolación trigonométrica, por lo cual los métodos directos para evaluar las constantes requieren millones de operaciones con multiplicaciones y sumas. El error de redondeo asociado a tal cantidad de cálculos generalmente domina la aproximación.

En 1965 un trabajo de J. W. Cooley y J. W. Tukey, publicado en la revista *Mathematics of Computation* [CT], describió otro método para calcular las constantes en el polinomio trigonométrico interpolante. El método requiere apenas $O(m \log_2 m)$ multiplicaciones y $O(m \log_2 m)$ sumas, siempre y cuando m se elija en forma correcta. En el caso de un problema con miles de puntos de datos, esto disminuye en miles la cantidad de cálculos, en comparación con los millones que se emplean en el método directo. Éste realmente fue descubierto varios años antes de publicarse el trabajo de Cooley-Tukey, pero había pasado inadvertido. ([Brigh, pp. 8-9], contiene un resumen breve e interesante del método.)

El método descrito por Cooley y Tukey se conoce con el nombre de **algoritmo de Cooley-Tukey** o bien **algoritmo de la transformada rápida de Fourier (TRF)** y ha provocado una verdadera revolución en la utilización de los polinomios trigonométricos interpolantes. Consiste en organizar el problema de manera que el número de puntos de datos a usar pueda factorizarse fácilmente, sobre todo en potencias de dos.

En vez de evaluar directamente las constantes a_k y b_k, la transformada rápida de Fourier calcula los coeficientes complejos c_k en

$$\frac{1}{m}\sum_{k=0}^{2m-1} c_k e^{ikx}, \tag{8.27}$$

donde

$$c_k = \sum_{k=0}^{2m-1} y_j e^{ik\pi j/m}, \qquad \text{para cada } k = 0, 1, \ldots, 2m-1. \tag{8.28}$$

Una vez que las constantes c_k se han determinado, a_k y b_k pueden recuperarse usando la *fórmula de Euler*,

$$e^{iz} = \cos z + i \operatorname{sen} z,$$

para cada $k = 0, 1, \ldots, m$,

$$\frac{1}{m}c_k(-1)^k = \frac{1}{m}c_k e^{-i\pi k} = \frac{1}{m}\sum_{j=0}^{2m-1} y_j e^{ik\pi j/m} e^{-i\pi k}\ \frac{1}{m}\sum_{j=0}^{2m-1} y_j e^{ik(-\pi+(\pi j/m))}$$

$$= \frac{1}{m}\sum_{j=0}^{2m-1} y_j\left(\cos k\left(-\pi + \frac{\pi j}{m}\right) + i \operatorname{sen} k\left(-\pi + \frac{\pi j}{m}\right)\right)$$

$$= \frac{1}{m}\sum_{j=0}^{2m-1} y_j (\cos kx_j + i \operatorname{sen} kx_j),$$

por tanto,

$$a_k + ib_k = \frac{(-1)^k}{m}c_k. \tag{8.29}$$

Con el fin de simplificar la notación, se agregan b_0 y b_m al conjunto, pero ambos son cero y por ello no influyen en la suma resultante.

La característica de reducción de operaciones de la transformada rápida de Fourier se debe al cálculo de los coeficientes c_k en conglomerados y utiliza como relación básica el hecho de que para todo entero n,

$$e^{n\pi i} = \cos n\pi + i \operatorname{sen} n\pi = (-1)^n.$$

Supongamos $m = 2^p$ para algún entero positivo p. Para cada $k = 0, 1, \ldots, m - 1$,

$$c_k + c_{m+k} = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m} + \sum_{j=0}^{2m-1} y_j e^{i(m+k)\pi j/m} = \sum_{j=0}^{2m-1} y_j e^{ik\pi j/m}(1 + e^{\pi ij}).$$

Pero

$$1 + e^{i\pi j} = \begin{cases} 2, & \text{si } j \text{ es par,} \\ 0, & \text{si } j \text{ es impar,} \end{cases}$$

en consecuencia, hay sólo m términos no ceros que deben sumarse. Si j se reemplaza con $2j$ en el índice de la suma, podemos escribir la suma como

$$c_k + c_{m+k} = 2\sum_{j=0}^{m-1} y_{2j} e^{ik\pi(2j)/m};$$

es decir,

$$c_k + c_{m+k} = 2\sum_{j=0}^{m-1} y_{2j} e^{ik\pi j/(m/2)}. \tag{8.30}$$

De manera semejante,

$$c_k - c_{m+k} = 2e^{ik\pi/m}\sum_{j=0}^{m-1} y_{2j+1} e^{ik\pi j/(m/2)}. \tag{8.31}$$

Puesto que c_k y c_{m+k} pueden recuperarse de las ecuaciones (8.30) y (8.31), estas relaciones determinan todos los coeficientes c_k. Nótese que las sumas de las ecuaciones (8.30) y (8.31) tienen la misma forma que la suma de la ecuación (8.28), salvo que el índice m ha sido sustituido por $m/2$.

Hay $2m$ coeficientes $c_0, c_1, \ldots, c_{2m-1}$ para calcular. El uso de la fórmula básica (8.28) requiere $2m$ multiplicaciones complejas por coeficiente, lo cual da un total de $(2m)^2$ operaciones. La ecuación (8.30) requiere m multiplicaciones complejas para cada $k = 0, 1, \ldots, m-1$, y (8.31) requiere $m + 1$ multiplicaciones complejas para cada $k = 0, 1, \ldots, m-1$. Al usar estas ecuaciones para calcular $c_0, c_1, \ldots, c_{2m-1}$ el número de multiplicaciones complejas disminuye de $(2m)^2 = 4m^2$ hasta

$$m \cdot m + m(m+1) = 2m^2 + m.$$

Como las sumas en (8.30) y (8.31) tienen la misma forma que la original y como m es una potencia de 2, podemos reaplicar el método de reducción a las sumas en (8.30) y (8.31). Éstas son reemplazadas por dos sumas de $j = 0$ a $j = (m/2) - 1$. Con ello la parte $2m^2$ de la suma se reduce a

$$2\left[\frac{m}{2} \cdot \frac{m}{2} + \frac{m}{2} \cdot \left(\frac{m}{2} + 1\right)\right] = m^2 + m.$$

Por tanto, ahora se necesitan un total de

$$(m^2 + m) + m = m^2 + 2m$$

multiplicaciones complejas.

Al aplicar el método una vez más, se obtienen 4 sumas, cada una con $m/4$ términos y la parte m^2 de este total se reduce a

$$4\left[\left(\frac{m}{4}\right)^2 + \frac{m}{4}\left(\frac{m}{4} + 1\right)\right] = \frac{m^2}{2} + m,$$

lo cual nos da un nuevo total de $(m^2/2) + 3m$ multiplicaciones complejas. Al repetir el proceso r veces, la cantidad total de multiplicaciones complejas necesarias se reduce a

$$\frac{m^2}{2^{r-2}} + mr.$$

El proceso concluye cuando $r = p + 1$ dado que $m = 2^p$ y $2m = 2^{p+1}$. En consecuencia, después de $r = p + 1$ reducciones de este tipo, el número de multiplicaciones complejas se reduce a

$$\frac{(2^p)^2}{2^{p-1}} + m(p+1) = 2m + pm + m = 3m + m\log_2 m = O(m\log_2 m).$$

Debido a la forma en que están arreglados los cálculos, el número de sumas complejas requeridas es semejante. Para explicar con un ejemplo la importancia de esta reducción, supóngase que tenemos $m = 2^{10} = 1024$. El calculo directo requeriría

$$(2m)^2 = (2048)^2 \approx 4\,200\,000.$$

El primer método de la transformada rápida de Fourier reduce la cantidad de cálculos a

$$3(1024) + 1024\log_2 1024 \approx 13\,300.$$

EJEMPLO 1 Consideremos el método de la transformada de Fourier aplicada a $8 = 2^3$ puntos de datos $\{(x_j, y_j)\}_{j=0}^{7}$ donde $x_j = -\pi + j\pi/4$, para cada $j = 0, 1, \ldots, 7$. En este caso $2m = 8$, así que $m = 4 = 2^2$ y $p = 2$.

De acuerdo con (8.24) tenemos

$$S_4(x) = \frac{a_0 + a_4 \cos 4x}{2} + \sum_{k=1}^{3} (a_k \cos kx + b_k \operatorname{sen} kx),$$

donde

$$a_k = \frac{1}{4} \sum_{j=0}^{7} y_j \cos kx_j \quad \text{y} \quad b_k = \frac{1}{4} \sum_{j=0}^{7} y_j \operatorname{sen} kx_j, \quad k = 0, 1, 2, 3, 4.$$

Definimos

$$F(x) = \frac{1}{4} \sum_{j=0}^{7} c_k e^{ikx},$$

donde

$$c_k = \sum_{j=0}^{7} y_j e^{ik\pi j/4}, \text{ para } k = 0, 1, \ldots, 7.$$

Entonces conforme a (8.29), para $k = 0, 1, 2, 3, 4$,

$$\frac{1}{4} c_k e^{-ik\pi} = a_k + ib_k.$$

Por cálculo directo, las constantes complejas c_k están dadas por

$$c_0 = y_0 + y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7;$$

$$c_1 = y_0 + ((i + 1)/\sqrt{2})\, y_1 + iy_2 + ((i - 1)/\sqrt{2})y_3 - y_4 - ((i + 1)/\sqrt{2})y_5 - iy_6 - ((i - 1)/\sqrt{2})y_7;$$

$$c_2 = y_0 + iy_1 - y_2 - iy_3 + y_4 + iy_5 - y_6 - iy_7;$$

$$c_3 = y_0 + ((i + 1)/\sqrt{2})y_1 - iy_2 + ((i + 1)/\sqrt{2})y_3 - y_4 - ((i - 1)/\sqrt{2})y_5 + iy_6 - ((i + 1)/\sqrt{2})y_7;$$

$$c_4 = y_0 - y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7;$$

$$c_5 = y_0 - ((i + 1)/\sqrt{2})y_1 + iy_2 - ((i - 1)/\sqrt{2})y_3 - y_4 + ((i + 1)/\sqrt{2})y_5 - iy_6 + ((i - 1)/\sqrt{2})y_7;$$

$$c_6 = y_0 - iy_1 - y_2 + iy_3 + y_4 - iy_5 - y_6 + iy_7;$$

$$c_7 = y_0 - ((i - 1)/\sqrt{2})y_1 - iy_2 - ((i + 1)/\sqrt{2})y_3 - y_4 + ((i - 1)/\sqrt{2})y_5 + iy_6 + ((i + 1)/\sqrt{2})y_7.$$

Debido al tamaño pequeño del conjunto de puntos de datos, muchos de los coeficientes de y_j en estas ecuaciones son 1 o -1. Esta frecuencia disminuirá en una aplicación más grande, por lo cual se incluirá la multiplicación por 1 o -1 para contar exactamente las operaciones de cálculo, aunque en este ejemplo no será necesario. Aclarado lo anterior, se requieren 64 multiplicaciones/divisiones y 56 sumas/restas en el cálculo directo de c_0, $c_1, \ldots, c_7$.

Al aplicar el procedimiento de la transformada rápida de Fourier con $r = 1$, primero definimos

$$d_0 = \frac{1}{2}(c_0 + c_4) = y_0 + y_2 + y_4 + y_6;$$

$$d_1 = \frac{1}{2}(c_0 - c_4) = y_1 + y_3 + y_5 + y_7;$$

$$d_2 = \frac{1}{2}(c_1 + c_5) = y_0 + iy_2 - y_4 - iy_6;$$

$$d_3 = \frac{1}{2}(c_1 - c_5) = ((i + 1)/\sqrt{2})(y_1 + iy_3 - y_5 - iy_7);$$

$$d_4 = \frac{1}{2}(c_2 + c_6) = y_0 - y_2 + y_4 - y_6;$$

$$d_5 = \frac{1}{2}(c_2 - c_6) = i(y_1 - y_3 + y_5 - y_7);$$

$$d_6 = \frac{1}{2}(c_3 + c_7) = y_0 - iy_2 - y_4 + iy_6;$$

$$d_7 = \frac{1}{2}(c_3 - c_7) = ((i - 1)/\sqrt{2})(y_1 - iy_3 - y_5 + iy_7).$$

Entonces definimos para $r = 2$.

$$e_0 = \frac{1}{2}(d_0 + d_4) = y_0 + y_4;$$

$$e_1 = \frac{1}{2}(d_0 - d_4) = y_2 + y_6;$$

$$e_2 = \frac{1}{2}(id_1 + d_5) = i(y_1 + y_5);$$

$$e_3 = \frac{1}{2}(id_1 - d_5) = i(y_3 + y_7);$$

$$e_4 = \frac{1}{2}(d_2 + d_6) = y_0 - y_4;$$

$$e_5 = \frac{1}{2}(d_2 - d_6) = i(y_2 - y_6);$$

$$e_6 = \frac{1}{2}(id_3 + d_7) = ((i - 1)/\sqrt{2})(y_1 - y_5);$$

$$e_7 = \frac{1}{2}(id_3 - d_7) = i((i - 1)/\sqrt{2})(y_3 - y_7).$$

Finalmente, para $r = p + 1 = 3$ definimos

$$f_0 = \frac{1}{2}(e_0 + e_4) = y_0;$$

$$f_1 = \frac{1}{2}(e_0 - e_4) = y_4;$$

$$f_2 = \frac{1}{2}(ie_1 + e_5) = iy_2;$$

$$f_3 = \frac{1}{2}(ie_1 - e_5) = iy_6;$$

$$f_4 = \frac{1}{2}(((i + 1)/\sqrt{2})e_2 + e_6) = ((i - 1)/\sqrt{2})y_1;$$

$$f_5 = \frac{1}{2}(((i + 1)/\sqrt{2})e_2 - e_6) = ((i - 1)/\sqrt{2})y_5;$$

$$f_6 = \frac{1}{2}(((i - 1)/\sqrt{2})e_3 + e_7) = (-(i + 1)/\sqrt{2})y_3;$$

$$f_7 = \frac{1}{2}(((i - 1)/\sqrt{2})e_3 - e_7) = (-(i + 1)/\sqrt{2})y_7.$$

Tanto $c_0, \ldots, c_7, d_0, \ldots, d_7, e_0, \ldots, e_7$, como $f_0, \ldots, f_7$ son independientes de los puntos particulares de datos; sólo dependen del hecho de que $m = 4$. Para toda m existe un conjunto único de constantes $\{c_k\}_{k=0}^{2m-1}$, $\{d_k\}_{k=0}^{2m-1}$, $\{e_k\}_{k=0}^{2m-1}$ y $\{f_k\}_{k=0}^{2m-1}$. Esta parte del trabajo no se necesita en una aplicación concreta. Sólo se requieren los siguientes cálculos:

1. $f_0 = y_0$; $\quad f_1 = y_4$; $\quad f_2 = iy_2$; $\quad f_3 = iy_6$;
   $f_4 = ((i-1)/\sqrt{2})y_1$; $\quad f_5 = ((i-1)/\sqrt{2})y_5$; $\quad f_6 = (-(i+1)/\sqrt{2})y_3$;
   $f_7 = (-(i+1)/\sqrt{2})y_7$.
2. $e_0 = f_0 + f_1$; $\quad e_1 = -i(f_2 + f_3)$; $\quad e_2 = ((-i+1)/\sqrt{2})(f_4 + f_5)$;
   $e_3 = ((-i-1)/\sqrt{2})(f_6 + f_7)$; $\quad e_4 = f_0 - f_1$; $\quad e_5 = f_2 - f_3$;
   $e_6 = f_4 - f_5$; $\quad e_7 = f_6 - f_7$.
3. $d_0 = e_0 + e_1$; $\quad d_1 = -i(e_2 + e_3)$; $\quad d_2 = e_4 + e_5$; $\quad d_3 = -i(e_6 + e_7)$;
   $d_4 = e_0 - e_1$; $\quad d_5 = e_2 - e_3$; $\quad d_6 = e_4 - e_5$; $\quad d_7 = e_6 - e_7$;
4. $c_0 = d_0 + d_1$; $\quad c_1 = d_2 + d_3$; $\quad c_2 = d_4 + d_5$; $\quad c_3 = d_6 + d_7$;
   $c_4 = d_0 - d_1$; $\quad c_5 = d_2 - d_3$; $\quad c_6 = d_4 - d_5$; $\quad c_7 = d_6 - d_7$.

Para calcular las constantes $c_0, c_1, \ldots, c_7$ en esta forma se requiere el número de operaciones que se muestra en la tabla 8.13. Nótese una vez más que la multiplicación por 1 o -1 ha sido incluida en el conteo, aunque para ello no hay que hacer cálculos.

La ausencia de multiplicaciones/divisiones en el paso 4 refleja el hecho de que, para cualquier m, los coeficientes $\{c_k\}_{k=0}^{2m-1}$ se calculan a partir de $\{d_k\}_{k=0}^{2m-1}$ en la misma forma:

$$c_k = d_{2k} + d_{2k+1}$$

Tabla 8.13

Paso	Multiplicaciones/divisiones	Sumas/restas
(1)	8	0
(2)	8	8
(3)	8	8
(4)	0	8
Total	24	24

y

$$c_{k+m} = d_{2k} - d_{2k+1}, \quad \text{para } k = 0, 1, \ldots, m-1,$$

de modo que no interviene ninguna multiplicación compleja.

En resumen, el cálculo directo de los coeficientes $c_0, c_1, \ldots, c_7$ requiere 64 multiplicaciones/divisiones y 56 sumas/restas. El método de la transformada rápida de Fourier reduce los cálculos a 24 multiplicaciones/divisiones y 24 sumas/restas. ■

El algoritmo 8.3 realiza la transformada rápida de Fourier cuando $m = 2^p$ para algún entero positivo p. Las modificaciones del método pueden hacerse cuando m adopta otras formas.

ALGORITMO 8.3

Método de la transformada rápida de Fourier

Para calcular los coeficientes de la suma

$$\frac{1}{m}\sum_{k=0}^{2m-1} c_k e^{ikx} = \frac{1}{m}\sum_{k=0}^{2m-1} c_k(\cos kx + i \operatorname{sen} kx), \quad \text{donde } i = \sqrt{-1},$$

para los datos $\{(x_j, y_j)\}_{j=0}^{2m-1}$ donde $m = 2^p$ y $x_j = -\pi + j\pi/m$ para $j = 0, 1, \ldots, 2m-1$:

ENTRADA $m, p; y_0, y_1, \ldots, y_{2m-1}$.

SALIDA números complejos $c_0, \ldots, c_{2m-1}$; números reales $a_0, \ldots, a_m; b_1, \ldots, b_{m-1}$.

Paso 1 Tome $M = m$;
$q = p$;
$\zeta = e^{\pi i/m}$.

Paso 2 Para $j = 0, 1, \ldots, 2m-1$ tome $c_j = y_j$.

Paso 3 Para $j = 1, 2, \ldots, M$ tome $\xi_j = \zeta^j$;
$\xi_{j+M} = -\xi_j$.

Paso 4 Tome $K = 0$;
$\xi_0 = 1$.

Paso 5 Para $L = 1, 2, \ldots, p + 1$ haga los pasos 6-12.

Paso 6 Mientras $K < 2m - 1$ haga los pasos 7-11.

Paso 7 Para $j = 1, 2, \ldots, M$ haga los pasos 8-10.

Paso 8 Sea $K = k_p \cdot 2^p + k_{p-1} \cdot 2^{p-1} + \cdots + k_1 \cdot 2 + k_0$;
(*Descomponga k.*)
tome $K_1 = K/2^q = k_p \cdot 2^{p-q} + \cdots + k_{q+1} \cdot 2 + k_q$;
$K_2 = k_q \cdot 2^p + k_{q+1} \cdot 2^{p-1} + \cdots + k_p \cdot 2^q$.

Paso 9 Tome $\eta = c_{K+M}\xi_{K_2}$;
$c_{K+M} = c_K - \eta$;
$c_K = c_K + \eta$.

Paso 10 Tome $K = K + 1$.

Paso 11 Tome $K = K + M$.

Paso 12 Tome $K = 0$;
$M = M/2$;
$q = q - 1$.

Paso 13 Mientras $K < 2m - 1$ haga los pasos 14-16.

Paso 14 Sea $K = k_p \cdot 2^p + k_{p-1} \cdot 2^{p-1} + \cdots + k_1 \cdot 2 + k_0$; (*Descomponga k.*)
tome $j = k_0 \cdot 2^p + k_1 \cdot 2^{p-1} + \cdots + k_{p-1} \cdot 2 + k_p$.

Paso 15 Si $j > K$ entonces intercambie c_j y c_k.

Paso 16 Tome $K = K + 1$.

Paso 17 Tome $a_0 = c_0/m$;
$a_m = \mathrm{Re}(e^{-i\pi m} c_m/m)$.

Paso 18 Para $j = 1, \ldots, m - 1$ tome $a_j = \mathrm{Re}(e^{-i\pi j} c_j/m)$;
$b_j = \mathrm{Im}(e^{-i\pi j} c_j/m)$.

Paso 19 SALIDA $(c_0, \ldots, c_{2m-1}; a_0, \ldots, a_m; b_1, \ldots, b_{m-1})$;
PARAR. ■

EJEMPLO 2 Sea $f(x) = x^4 - 3x^3 + 2x^2 - \tan x(x - 2)$. Para determinar el polinomio trigonométrico interpolante de cuarto grado para los datos $\{(x_j, y_j)\}_{j=0}^{7}$, donde $x_j = j/4$ y $y_j = f(x_j)$, es necesario transformar el intervalo $[0, 2]$ a $[-\pi, \pi]$. La traslación lineal está dada por

$$z_j = \pi(x_j - 1),$$

así que los datos de entrada del algoritmo 8.3 son

$$\left\{z_j, f\left(1 + \frac{z_j}{\pi}\right)\right\}_{j=0}^{7}.$$

El polinomio interpolante en z es

$$S_4(z) = 0.761979 + 0.771841 \cos z + 0.0173037 \cos 2z + 0.00686304 \cos 3z$$
$$-0.000578545 \cos 4z - 0.386374 \operatorname{sen} z + 0.0468750 \operatorname{sen} 2z - 0.0113738 \operatorname{sen} 3z.$$

El polinomio trigonométrico $S_4(x)$ en $[0, 2]$ se obtiene sustituyendo $z = \pi(x - 1)$ en $S_4(z)$. Las gráficas de $y = f(x)$ y $y = S_4(x)$ se muestran en la figura 8.15 y en la tabla 8.14 se incluyen los valores de $f(x)$ y $S_4(x)$. ■

Figura 8.15

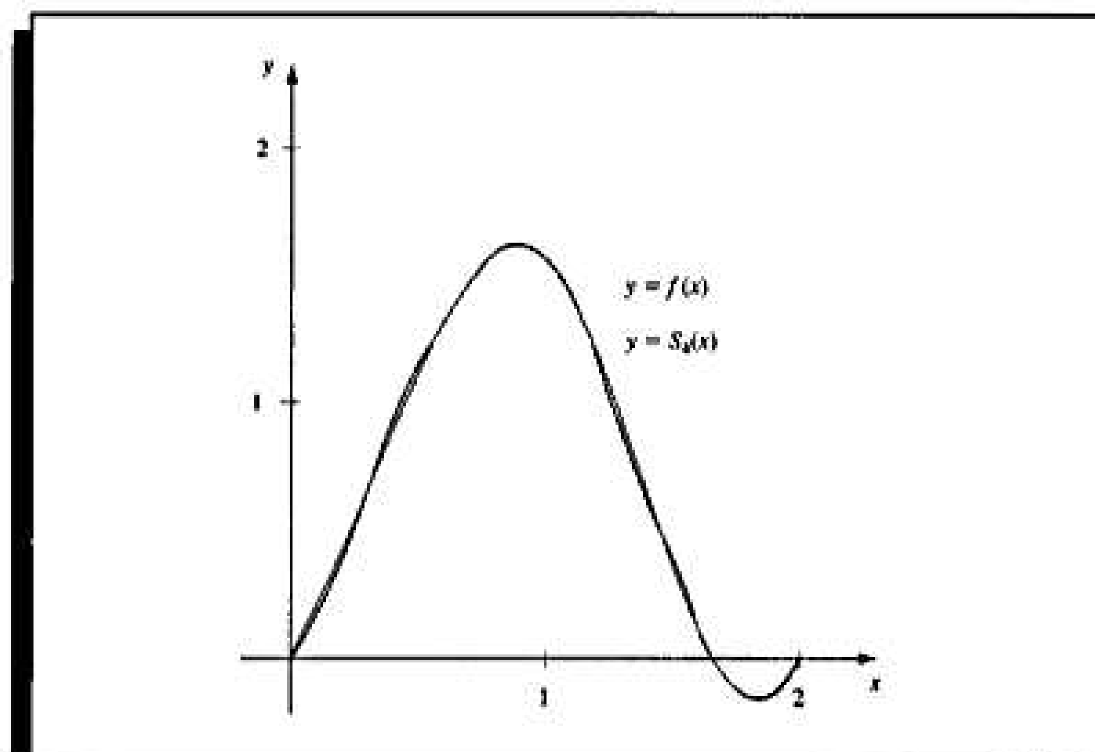


Tabla 8.14

x	$f(x)$	$S_4(x)$	$\|f(x) - S_4(x)\|$
0.125	0.26440	0.25001	1.44×10^{-2}
0.375	0.84081	0.84647	5.66×10^{-3}
0.625	1.36150	1.35824	3.27×10^{-3}
0.875	1.61282	1.61515	2.33×10^{-3}
1.125	1.36672	1.36471	2.02×10^{-3}
1.375	0.71697	0.71931	2.33×10^{-3}
1.625	0.07909	0.07496	4.14×10^{-3}
1.875	−0.14576	−0.13301	1.27×10^{-2}

El lector que desee profundizar en la comprobación de la validez del método de la transformada rápida de Fourier, puede consultar a [Ham], quien presenta el método desde un enfoque matemático, o a [Bra], cuya presentación se basa en métodos con los que están más familiarizados los ingenieros. [AHU, pp. 252-269] es una buena obra de consulta respecto a los aspectos computacionales del método. En [Win] se describe la modificación del procedimiento en el caso en que m no es una potencia de 2. En [Lau, pp. 438-465] se incluye una explicación de la técnica y de temas conexos desde el punto de vista del álgebra abstracta aplicada.

CONJUNTO DE EJERCICIOS 8.6

1. Determine el polinomio trigonométrico interpolante $S_2(x)$ de segundo grado en $[-\pi, \pi]$ para las siguientes funciones, y después grafique $f(x) - S_2(x)$:

a. $f(x) = \pi(x - \pi)$ **b.** $f(x) = x(\pi - x)$

c. $f(x) = |x|$ **d.** $f(x) = \begin{cases} -1, & -\pi \le x \le 0 \\ 1, & 0 < x \le \pi \end{cases}$

2. Determine el polinomio trigonométrico interpolante de cuarto grado para $f(x) = x(\pi - x)$ en el intervalo $[-\pi, \pi]$ por medio de

a. El cálculo directo;

b. El algoritmo de la transformada rápida de Fourier.

3. Use el algoritmo de la transformada rápida de Fourier para calcular el polinomio trigonométrico interpolante de cuarto grado en $[-\pi, \pi]$ para las siguientes funciones.

a. $f(x) = \pi(x - \pi)$ **b.** $f(x) = |x|$

c. $f(x) = \cos \pi x - 2 \operatorname{sen} \pi x$ **d.** $f(x) = x \cos x^2 + e^x \cos e^x$

4. **a.** Determine el polinomio trigonométrico interpolante $S_4(x)$ de cuarto grado para $f(x) = x^2 \operatorname{sen} x$ en el intervalo $[0, 1]$.

b. Calcule $\int_0^1 S_4(x)\, dx$.

c. Compare la integral de la parte (b) con $\int_0^1 x^2 \operatorname{sen} x\, dx$.

5. Use las aproximaciones obtenidas en el ejercicio 3 para aproximar las siguientes integrales, y después compare los resultados con los valores reales.

a. $\int_{-\pi}^{\pi} \pi(x - \pi)\, dx$ **b.** $\int_{-\pi}^{\pi} |x| dx$

c. $\int_{-\pi}^{\pi} (\cos \pi x - 2 \operatorname{sen} \pi x)\, dx$ **d.** $\int_{-\pi}^{\pi} (x \cos x^2 + e^x \cos e^x)\, dx$

6. Use el algoritmo de la transformada rápida de Fourier para determinar el polinomio trigonométrico interpolante de decimosexto grado para $f(x) = x^2 \cos x$ en $[-\pi, \pi]$.

7. Use el algoritmo de la transformada rápida de Fourier para determinar el polinomio trigonométrico interpolante de sexagésimo cuarto grado para $f(x) = x^2 \cos x$ en $[-\pi, \pi]$.

8. Use la identidad trigonométrica para demostrar que $\sum_{j=0}^{2m-1} (\cos mx_j)^2 = 2m$.

9. Demuestre que $c_0, \ldots, c_{2m-1}$ en el algoritmo 8.3 están dados por

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{2m-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \zeta & \zeta^2 & \cdots & \zeta^{2m-1} \\ 1 & \zeta^2 & \zeta^4 & \cdots & \zeta^{4m-2} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \zeta^{2m-1} & \zeta^{4m-2} & \cdots & \zeta^{(2m-1)^2} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{2m-1} \end{bmatrix},$$

donde $\zeta = e^{\pi i/m}$.

10. En la exposición que se hizo antes del algoritmo 8.3, se explicó un ejemplo para $m = 4$. Defina los vectores **c**, **d**, **e**, **f** y **y** como

$$\mathbf{c} = (c_0, c_1, \ldots, c_7)^t,$$
$$\mathbf{d} = (d_0, d_1, \ldots, d_7)^t,$$
$$\mathbf{e} = (e_0, e_1, \ldots, e_7)^t,$$

$$\mathbf{f} = (f_0, f_1, \ldots, f_7)^t,$$
$$\mathbf{y} = (y_0, y_1, \ldots, y_7)^t.$$

Obtenga las matrices A, B, C y D de modo que $\mathbf{c} = A\mathbf{d}$, $\mathbf{d} = B\mathbf{e}$, $\mathbf{e} = C\mathbf{f}$ y $\mathbf{f} = D\mathbf{y}$.

8.7 Reseña de métodos y de software

En el presente capítulo estudiamos la aproximación de datos y funciones por medio de funciones elementales. Las funciones elementales que se usaron fueron polinomios, funciones racionales y polinomios trigonométricos. También explicamos dos tipos de aproximaciones: las discretas y las continuas. Las primeras ocurren cuando aproximamos un conjunto finito de datos mediante una función elemental. Las segundas se emplean cuando se conoce la función por aproximar.

El uso de los métodos de mínimos cuadrados discretos se recomienda cuando la función se especifica mediante un conjunto de datos que no necesariamente representan la función. El ajuste de mínimos cuadrados para los datos puede adoptar la forma de una aproximación polinomial lineal o de otro tipo, e incluso una forma exponencial. Estas aproximaciones se calculan resolviendo conjuntos de ecuaciones normales, como las de la sección 8.1.

Si los datos son periódicos, posiblemente convenga utilizar el ajuste trigonométrico de mínimos cuadrados. Debido a la ortogonalidad de las funciones trigonométricas básicas, en la aproximación trigonométrica mediante mínimos cuadrados no es necesario resolver un sistema lineal. Con grandes cantidades de datos periódicos, también se recomienda la interpolación por medio de polinomios trigonométricos. Una forma eficiente de calcular el polinomio trigonométrico interpolante es la transformada rápida de Fourier.

Cuando la función que va a aproximarse se puede evaluar en cualquier argumento requerido, las aproximaciones tratan de reducir al mínimo la integral en vez de la suma. En la sección 8.2 se estudiaron las aproximaciones polinomiales de mínimos cuadrados continuos. El cálculo eficiente de los polinomios de mínimos cuadrados da origen a conjuntos ortogonales de polinomios, como los de Legendre y los de Chebyshev. En la sección 8.4 se estudió la aproximación mediante funciones racionales, donde explicamos la aproximación de Padé como una generalización del polinomio de Maclaurin, y también explicamos la extensión a la aproximación racional de Chebyshev. Ambos métodos son procedimientos de aproximación más uniformes que los polinomios. En la sección 8.5 se examinó la aproximación de los mínimos cuadrados continuos por medio de funciones trigonométricas, especialmente en lo concerniente a las series de Fourier.

La biblioteca IMSL ofrece varias rutinas de aproximación. La subrutina RLINE da la línea de mínimos cuadrados para un conjunto de puntos de datos, y produce estadísticas como las medias y las varianzas. La subrutina FNLSQ calcula la aproximación de mínimos cuadrados discretos según la elección de las funciones básicas por parte del usuario, y BSLSQ calcula una aproximación de mínimos cuadrados para los trazadores cúbicos. La subrutina RATCH calcula la aproximación racional ponderada de Chebyshev a una función continua en el intervalo $[a, b]$. La subrutina FFTCB calcula la transformada rápida de Fourier para determinado conjunto de datos, y se parece al algoritmo 8.3.

La biblioteca NAG tiene muchas subrutinas para aproximar funciones. La aproximación polinomial de mínimos cuadrados se obtiene en la subrutina E02ADF. Esta subrutina es muy flexible, ya que calcula los polinomios de mínimos cuadrados para diversos gra-

dos y proporciona sus errores de mínimos cuadrados. Utiliza los polinomios de Chebyshev para reducir al mínimo el error de redondeo y aumentar la exactitud.

La rutina E02AEF se usa para evaluar la aproximación obtenida mediante E02ADF. NAG también ofrece la rutina E02BAF para calcular los ajustes de mínimos cuadrados para los trazadores cúbicos, E02GAF calcula el mejor ajuste lineal L_1 y E02GCF calcula el mejor ajuste L_∞. La rutina E02RAF calcula la aproximación de Padé. La biblioteca NAG también contiene muchas rutinas para las transformadas rápidas de Fourier, una de las cuales es C06ECF.

La biblioteca netlib contiene la subrutina polfit.f dentro del paquete slatec para calcular la aproximación polinomial por mínimos cuadrados para un conjunto discreto de puntos. La subrutina pvalue.f se puede usar para evaluar el polinomio obtenido con polfit.f (o cualquiera de sus derivadas en un punto dado).

El lector que desee información más completa sobre la teoría general de la aproximación puede consultar a Powell [Po], a Davis [Da] o a Cheney [Ch]. Un buen libro sobre los métodos de los mínimos cuadrados es el de Lawson y Hanson [LH]; en Van Loan [Van] y en Briggs y Hanson [BH] puede encontrar información acerca de las transformadas de Fourier.

CAPÍTULO

9 Aproximación de los valores característicos

• • •

Las vibraciones longitudinales de una barra elástica de rigidez $p(x)$ y densidad $\rho(x)$ locales se describen mediante la ecuación diferencial parcial

$$\rho(x)\frac{\partial^2 v}{\partial t^2}(x,t) = \frac{\partial}{\partial x}\left[p(x)\frac{\partial v}{\partial x}(x,t)\right],$$

donde $v(x, t)$ es el desplazamiento promedio longitudinal de una sección de la barra, desde su posición de equilibrio x en el tiempo t. Podemos escribir las vibraciones como una suma de vibraciones armónicas simples:

$$v(x,t) = \sum_{k=0}^{\infty} c_k u_k(x) \cos \sqrt{\lambda_k}(t - t_0),$$

donde

$$\frac{d}{dx}\left[p(x)\frac{du_k}{dx}(x)\right] + \lambda_k \rho(x) u_k(x) = 0.$$

Si la barra tiene una longitud l y está fija en sus extremos, entonces esta ecuación diferencial es válida para $0 < x < l$ y $v(0) = v(l) = 0$. Un sistema con estas ecuaciones diferenciales se denomina sistema de Sturm-Liouville, y los números λ_k son valores característicos con las correspondientes funciones características $u_k(x)$.

Supongamos que la barra tiene 1 m de longitud, una rigidez uniforme $p(x) = p$ y una densidad también uniforme $\rho(x) = \rho$. Para aproximar u y λ, sea $h = 0.2$. Entonces, $x_j = 0.2j$, para $0 \leq j \leq 5$, y podemos aplicar la fórmula de diferencia centrada (4.5) de la sección 4.1 para aproximar las primeras derivadas. Esto nos da el sistema lineal

$$A\mathbf{w} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = -0.04\frac{\rho}{p}\lambda\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = -0.04\frac{\rho}{p}\lambda\mathbf{w}.$$

En este sistema, $w_j = u(x_j)$, para $1 \leq j \leq 4$, y $w_0 = w_5 = 0$. Los cuatro valores característicos de A se aproximan a los del *sistema Sturm-Liouville*. La aproximación de los valores característicos es la que estudiaremos en este capítulo. En el ejercicio 11 de la sección 9.4 se estudia una aplicación de Sturm-Liouville.

9.1 Álgebra lineal y valores característicos

En el capítulo 7 presentamos los valores y vectores característicos en relación con la convergencia de métodos iterativos para aproximar la solución de un sistema lineal. Para determinar los valores característicos de una matriz A de $n \times n$, construimos el polinomio característico

$$p(\lambda) = \det(A - \lambda I)$$

y luego calculamos sus ceros. Desde el punto de vista de los cálculos, es tedioso obtener el determinante de una matriz $n \times n$ y también es difícil hallar buenas aproximaciones de las raíces de $p(\lambda)$. En este capítulo analizaremos otras formas de aproximar los valores característicos de una matriz.

En el capítulo 7 vimos que una técnica iterativa para resolver un sistema lineal convergerá si todos los valores característicos asociados con el problema tienen magnitud menor que 1. En este caso, los valores exactos de los valores característicos no tienen una importancia básica, sólo la región del plano complejo donde se encuentren.

Aunque necesitamos conocer los valores característicos, el hecho de que muchas de las técnicas utilizadas para aproximarlos sean iterativas implica que la determinación de la región donde se encuentran es un primer paso para determinar la aproximación, pues esto proporciona la aproximación inicial necesaria para los métodos iterativos.

Antes de estudiar otros resultados concernientes a los valores y vectores característicos, necesitamos recordar algunas definiciones y resultados del álgebra lineal. Todos los resultados generales que se requerirán en el resto del capítulo se incluyen aquí para facilitar su consulta. Las pruebas de los resultados que no se incluyen se encuentran prácticamente en cualquier libro de álgebra lineal (véase, por ejemplo, [ND]). La primera definición nos recuerda la de la independencia lineal de las funciones que se vio en la sección 8.2.

Definición 9.1 Sea $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(k)}\}$ un conjunto de vectores. El conjunto es **linealmente independiente**, en tanto

$$\mathbf{0} = \alpha_1\mathbf{v}^{(1)} + \alpha_2\mathbf{v}^{(2)} + \alpha_3\mathbf{v}^{(3)} + \cdots + \alpha_k\mathbf{v}^{(k)},$$

entonces, $\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_k = 0$. De lo contrario, el conjunto de vectores es **linealmente dependiente**. ■

Nótese que cualquier conjunto de vectores que contenga el vector cero es linealmente dependiente.

Teorema 9.2 Si $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(n)}\}$ es un conjunto de n vectores linealmente independientes de $\mathbb{R}^n$, entonces cualquier vector $\mathbf{x} \in \mathbb{R}^n$ puede escribirse de manera única como

$$\mathbf{x} = \beta_1\mathbf{v}^{(1)} + \beta_2\mathbf{v}^{(2)} + \beta_3\mathbf{v}^{(3)} + \cdots + \beta_n\mathbf{v}^{(n)},$$

para algunas constantes $\beta_1, \beta_2, \ldots, \beta_n$. ■

Demostración Supóngase que A es la matriz cuyas columnas son los vectores $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$. Entonces, el conjunto $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ es linealmente independiente si y sólo si la ecuación matricial $A\alpha = \mathbf{0}$ tiene la solución única $\alpha = \mathbf{0}$. Pero, de acuerdo con el teorema 6.16, esto equivale a afirmar que para cualquier vector $\mathbf{x} \in \mathbb{R}^n$, la ecuación matricial $A\beta = \mathbf{x}$ tiene una solución única. Y, a su vez, esto equivale a la afirmación de que para cualquier $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x} = \beta_1\mathbf{v}^{(1)} + \beta_2\mathbf{v}^{(2)} + \cdots + \beta_n\mathbf{v}^{(n)}$$

para algún conjunto único de constantes $\beta_1, \beta_2, \ldots, \beta_n$. ■ ■ ■

Todo conjunto de n vectores linealmente independientes en $\mathbb{R}^n$ recibe el nombre de base de $\mathbb{R}^n$.

EJEMPLO 1 Sean $\mathbf{v}^{(1)} = (1, 0, 0)^t$, $\mathbf{v}^{(2)} = (-1, 1, 1)^t$ y $\mathbf{v}^{(3)} = (0, 4, 2)^t$. Si α_1, α_2 y α_3 son números con

$$\mathbf{0} = \alpha_1\mathbf{v}^{(1)} + \alpha_2\mathbf{v}^{(2)} + \alpha_3\mathbf{v}^{(3)},$$

entonces

$$\begin{aligned}(0, 0, 0)^t &= \alpha_1(1, 0, 0)^t + \alpha_2(-1, 1, 1)^t + \alpha_3(0, 4, 2)^t \\ &= (\alpha_1 - \alpha_2, \alpha_2 + 4\alpha_3, \alpha_2 + 2\alpha_3)^t,\end{aligned}$$

así que

$$\alpha_1 - \alpha_2 = 0, \qquad \alpha_2 + 4\alpha_3 = 0 \qquad \text{y} \qquad \alpha_2 + 2\alpha_3 = 0.$$

La única solución de este sistema es $\alpha_1 = \alpha_2 = \alpha_3 = 0$ y, por tanto, el conjunto $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}$ es linealmente independiente en $\mathbb{R}^3$ y es base de $\mathbb{R}^3$.

Todo vector $\mathbf{x} = (x_1, x_2, x_3)^t$ en $\mathbb{R}^3$ puede escribirse como

$$\mathbf{x} = \beta_1\mathbf{v}^{(1)} + \beta_2\mathbf{v}^{(2)} + \beta_3\mathbf{v}^{(3)},$$

al seleccionar

$$\beta_1 = x_1 - x_2 + 2x_3, \qquad \beta_2 = 2x_3 - x_2 \qquad \text{y} \qquad \beta_3 = \frac{1}{2}(x_2 - x_3).$$ ■

El siguiente resultado se usará en la sección 9.2 para desarrollar el método de potencias y para aproximar los valores característicos. En el ejercicio 8 se incluye una demostración del resultado.

Teorema 9.3 Si A es una matriz y $\lambda_1, \ldots, \lambda_k$ son valores característicos distintos de A con los vectores característicos asociados $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}$, entonces $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}\}$ es linealmente independiente. ■

En la sección 8.2 estudiamos los conjuntos ortogonales y ortonormales de las funciones. De manera similar se definen los vectores que tienen esas propiedades.

Definición 9.4 Se dice que un conjunto de vectores $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ es **ortogonal** si $(\mathbf{v}^{(i)})^t\mathbf{v}^{(j)} = 0$ para toda $i \neq j$. Además, si $(\mathbf{v}^{(i)})^t\mathbf{v}^{(i)} = 1$ para toda $i = 1, 2, \ldots, n$, entonces se dice que el conjunto es **ortonormal**. ■

Puesto que $\mathbf{x}^t\mathbf{x} = \|\mathbf{x}\|_2^2$, un conjunto de vectores ortogonales $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ es ortonormal si y sólo si

$$\|\mathbf{v}^{(i)}\|_2 = 1, \qquad \text{para cada } i = 1, 2, \ldots, n.$$

EJEMPLO 2 Los vectores $\mathbf{v}^{(1)} = (0, 4, 2)^t$, $\mathbf{v}^{(2)} = (-1, -\frac{1}{5}, \frac{2}{5})^t$ y $\mathbf{v}^{(3)} = (\frac{1}{6}, -\frac{1}{6}, \frac{1}{3})^t$ forman un conjunto ortogonal. Las normas l_2 de ellos son

$$\|\mathbf{v}^{(1)}\|_2 = 2\sqrt{5}, \quad \|\mathbf{v}^{(2)}\|_2 = \frac{\sqrt{30}}{5} \quad \text{y} \quad \|\mathbf{v}^{(3)}\|_2 = \frac{\sqrt{6}}{6}.$$

En consecuencia, los vectores

$$\mathbf{u}^{(1)} = \frac{\mathbf{v}^{(1)}}{\|\mathbf{v}^{(1)}\|_2} = \left(0, \frac{2\sqrt{5}}{5}, \frac{\sqrt{5}}{5}\right)^t,$$

$$\mathbf{u}^{(2)} = \frac{\mathbf{v}^{(2)}}{\|\mathbf{v}^{(2)}\|_2} = \left(-\frac{\sqrt{30}}{6}, -\frac{\sqrt{30}}{30}, \frac{\sqrt{30}}{15}\right)^t,$$

y

$$\mathbf{u}^{(3)} = \frac{\mathbf{v}^{(3)}}{\|\mathbf{v}^{(3)}\|_2} = \left(\frac{\sqrt{6}}{6}, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3}\right)^t$$

forman un conjunto ortonormal, porque heredan la ortogonalidad de $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}$ y $\mathbf{v}^{(3)}$, además de que

$$\|\mathbf{u}^{(1)}\|_2 = \|\mathbf{u}^{(2)}\|_2 = \|\mathbf{u}^{(3)}\|_2 = 1.$$ ■

La comprobación del siguiente resultado se puede ver en el ejercicio 5.

Teorema 9.5 Un conjunto ortogonal de vectores distintos de cero es linealmente independiente. ■

La terminología de la siguiente definición proviene del hecho de que las columnas de una matriz ortogonal forman un conjunto de vectores ortogonales, más bien ortonormales. (Véase el ejercicio 6.)

Definición 9.6 Se dice que una matriz Q es **ortogonal** si $Q^{-1} = Q^t$. ■

Recuerde que las matrices de permutación que estudiamos en la sección 6.5 tienen esta propiedad y, por consiguiente, son ortogonales.

EJEMPLO 3 La matriz ortogonal Q formada a partir del conjunto ortonormal de vectores incluidos en el ejemplo 2 es

$$Q = [\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}] = \begin{bmatrix} 0 & -\frac{\sqrt{30}}{6} & \frac{\sqrt{6}}{6} \\ \frac{2\sqrt{5}}{5} & -\frac{\sqrt{30}}{30} & -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{5}}{5} & \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} \end{bmatrix}.$$

Nótese que

$$QQ^t = \begin{bmatrix} 0 & \frac{-\sqrt{30}}{6} & \frac{\sqrt{6}}{6} \\ \frac{2\sqrt{5}}{5} & -\frac{\sqrt{30}}{30} & -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{5}}{5} & \frac{\sqrt{30}}{15} & \frac{\sqrt{6}}{3} \end{bmatrix} \cdot \begin{bmatrix} 0 & \frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ -\frac{\sqrt{30}}{6} & -\frac{\sqrt{30}}{30} & \frac{\sqrt{30}}{15} \\ \frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

También es verdad que $Q^t Q = I$, por lo que $Q^t = Q^{-1}$. ■

Definición 9.7 Se dice que dos matrices A y B son **similares** si existe una matriz S no singular con $A = S^{-1}BS$. ■

El aspecto importante de las matrices similares es que tienen los mismos valores característicos.

Teorema 9.8 Supongamos que A y B son matrices similares con $A = S^{-1}BS$ y que λ es un valor característico de A con el vector característico asociado $\mathbf{x}$. Entonces, λ es un valor característico de B con $S\mathbf{x}$ como vector característico asociado. ■

Demostración Supongamos que $\mathbf{x} \neq \mathbf{0}$ es tal que

$$S^{-1}BS\mathbf{x} = A\mathbf{x} = \lambda\mathbf{x}.$$

Al multiplicar en la izquierda, por la matriz S, obtenemos

$$BS\mathbf{x} = \lambda S\mathbf{x}.$$

Puesto que $\mathbf{x} \neq \mathbf{0}$ y S es no singular, $S\mathbf{x} \neq \mathbf{0}$. Por tanto, $S\mathbf{x}$ es un vector característico de B asociado al valor característico λ. ■ ■ ■

El comando de Maple `issimilar (A, B)` regresa *verdadero* si A y B son similares, y *falso* en caso contrario.

Es fácil determinar los valores característicos de una matriz triangular A, porque en este caso λ es una solución de la ecuación

$$0 = \det(A - \lambda I) = \prod_{i=1}^{n}(a_{ii} - \lambda)$$

si y sólo si $\lambda = a_{ii}$ para alguna i. El siguiente resultado describe una relación, denominada **transformación de similitud**, entre matrices arbitrarias y triangulares.

Teorema 9.9 **(Teorema de Schur)**

Sea A una matriz arbitraria. Existe una matriz no singular U con la propiedad de que

$$T = U^{-1}AU,$$

donde T es una matriz triangular superior cuyos elementos diagonales constan de los valores característicos de A. ■

La matriz U, cuya existencia se garantiza con el teorema 9.9, satisface la condición $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ para cualquier vector $\mathbf{x}$. A este tipo de matrices se les llama **unitarias**. Aunque no usaremos esta propiedad de conservación de la norma, sí aumenta considerablemente la aplicación del teorema.

El teorema 9.9 es un teorema de existencia, pues garantiza la existencia de la matriz triangular T sin ofrecer un medio constructivo para obtener T. Para demostrar el teorema es necesario conocer los valores característicos de A. Así, en la mayoría de los casos es difícil determinar la transformación de similitud U. La siguiente restricción del teorema 9.9 en las matrices simétricas disminuye la complejidad, porque en este caso la matriz de transformación es ortogonal.

Teorema 9.10 Si A es una matriz simétrica y D es una matriz ortogonal cuyos elementos diagonales son los valores característicos de A, entonces existe una matriz ortogonal Q tal que $D = Q^{-1}AQ = Q^tAQ$. ■

Los siguientes corolarios del teorema 9.10 demuestran algunas de las propiedades más interesantes de las matrices simétricas.

Corolario 9.11 Si A es una matriz simétrica de $n \times n$, entonces los valores característicos de A son números reales, y existen n vectores característicos de A que forman un conjunto ortonormal. ■

Demostración Si $Q = (q_{ij})$ y $D = (d_{ij})$ son las matrices especificadas en el teorema 9.10, entonces

$$D = Q^{-1}AQ \quad \text{implica que} \quad AQ = QD.$$

Sea $1 \le i \le n$ y $\mathbf{v}_i = (q_{1i}, q_{2i}, \ldots, q_{ni})^t$ la i-ésima columna de Q. Entonces

$$A\mathbf{v}_i = d_{ii}\mathbf{v}_i,$$

y d_{ii} es un valor característico de A con vector característico $\mathbf{v}_i$, la i-ésima columna de Q. Como las columnas de Q son ortonormales, los vectores característicos de A también lo son. Al multiplicar esta ecuación a la izquierda por $\mathbf{v}_i^t$ se tiene

$$\mathbf{v}_i^tA\mathbf{v}_i = d_{ii}\mathbf{v}_i^t\mathbf{v}_i.$$

Como $\mathbf{v}_i^tA\mathbf{v}_i$ y $\mathbf{v}_i^t\mathbf{v}_i$ son números reales y $\mathbf{v}_i^t\mathbf{v}_i = 1$, el valor característico $d_{ii} = \mathbf{v}_i^tA\mathbf{v}_i$ es un número real, para cada $i = 1, 2, \ldots, n$. ■ ■ ■

En la sección 6.6 vimos que una matriz simétrica A es definida positiva si para todo vector $\mathbf{x}$ distinto de cero tenemos $\mathbf{x}^tA\mathbf{x} > 0$. El siguiente teorema describe las matrices definidas positivas a partir de los valores característicos. Esta propiedad de sus valores característicos las hace importantes en las aplicaciones.

Teorema 9.12 Una matriz simétrica A es definida positiva si y sólo si todos los valores característicos de A son positivos. ■

Demostración Supongamos primero que A es definida positiva y que λ es un valor característico de A con vector característico asociado $\mathbf{x}$. Entonces

$$0 < \mathbf{x}^t A\mathbf{x} = \lambda \mathbf{x}^t\mathbf{x} = \lambda\|\mathbf{x}\|_2^2,$$

por lo que $\lambda > 0$. En consecuencia, todo valor característico de una matriz definida positiva es positivo.

Para demostrar el recíproco, supóngase que A es simétrica con valores característicos positivos. De acuerdo con el corolario 9.11, existen n vectores característicos de A, $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$, que forman un conjunto ortonormal y, según el teorema 9.5, linealmente independiente. Así, para cualquier vector $\mathbf{x} \neq 0$ existe un conjunto único de constantes β_1, $\beta_2, \ldots, \beta_n$ distintas de cero y para las cuales

$$\mathbf{x} = \sum_{i=1}^{n} \beta_i \mathbf{v}^{(i)}.$$

Al multiplicar por $\mathbf{x}^t A$ obtenemos

$$\mathbf{x}^t A\mathbf{x} = \mathbf{x}^t\left(\sum_{i=1}^{n} \beta_i A\mathbf{v}^{(i)}\right) = \mathbf{x}^t\left(\sum_{i=1}^{n} \beta_i\lambda_i\mathbf{v}^{(i)}\right) = \sum_{j=1}^{n}\sum_{i=1}^{n} \beta_j\beta_i\lambda_i(\mathbf{v}^{(j)})^t\mathbf{v}^{(i)}.$$

Pero los vectores $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$ forman un conjunto ortonormal y, por tanto,

$$(\mathbf{v}^{(j)})^t\mathbf{v}^{(i)} = \begin{cases} 0, & \text{si } i \neq j, \\ 1, & \text{si } i = j. \end{cases}$$

Esto, con el hecho de que λ_i son todas positivas, implica que

$$\mathbf{x}^t A\mathbf{x} = \sum_{j=1}^{n}\sum_{i=1}^{n} \beta_j\beta_i\lambda_i(\mathbf{v}^{(j)})^t\mathbf{v}^{(i)} = \sum_{i=1}^{n} \lambda_i\beta_i^2 > 0.$$

Por tanto, A es definida positiva. ■ ■ ■

El resultado final de la sección se refiere a las cotas para aproximar los valores característicos.

Teorema 9.13 **(Teorema del círculo de Gerschgorin)**

Sea A una matriz de $n \times n$ y denotemos con R_i el círculo en el plano complejo con centro a_{ii} y radio $\sum_{j=1, j\neq i}^{n} |a_{ij}|$; es decir,

$$R_i = \left\{ z \in \mathcal{C} \,\middle|\, |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\},$$

donde $\mathcal{C}$ denota el plano complejo. Los valores característicos de A están contenidos dentro de $R = \cup_{i=1}^{n} R_i$. Más aún, la unión de cualquier k de estos círculos que no tenga una intersección con los $(n - k)$ restantes contendrá exactamente k (contando multiplicidades) de los valores característicos. ■

Demostración Supongamos que λ es un valor característico de A con el vector característico asociado $\mathbf{x}$, donde $\|\mathbf{x}\|_\infty = 1$. Dado que $A\mathbf{x} = \lambda\mathbf{x}$, la representación equivalente por componente es

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i, \qquad \text{para cada } i = 1, 2, \ldots, n.$$

Si k es un entero con $|x_k| = \|\mathbf{x}\|_\infty = 1$, esta ecuación, con $i = k$, implica que

$$\sum_{j=1}^{n} a_{kj}x_j = \lambda x_k.$$

Por tanto,

$$\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j = \lambda x_k - a_{kk}x_k = (\lambda - a_{kk})x_k.$$

y

$$|\lambda - a_{kk}| \cdot |x_k| = \left|\sum_{\substack{j=1,\\ j\neq k}}^{n} a_{kj}x_j\right| \leq \sum_{\substack{j=1,\\ j\neq k}}^{n} |a_{kj}||x_j|.$$

Como $|x_j| \leq |x_k| = 1$, para toda $j = 1, 2, \ldots, n$,

$$|\lambda - a_{kk}| \leq \sum_{\substack{j=1,\\ j\neq k}}^{n} |a_{kj}|.$$

Por tanto, $\lambda \in R_k$, lo cual prueba la primera afirmación del teorema. La segunda parte requiere un argumento de continuidad ingenioso. Una demostración explícita se encuentra en [Or2, p. 48]. ■ ■ ■

EJEMPLO 4 Para la matriz

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{bmatrix},$$

los círculos del teorema de Gerschgorin son (véase la figura 9.1 en la p. 558)

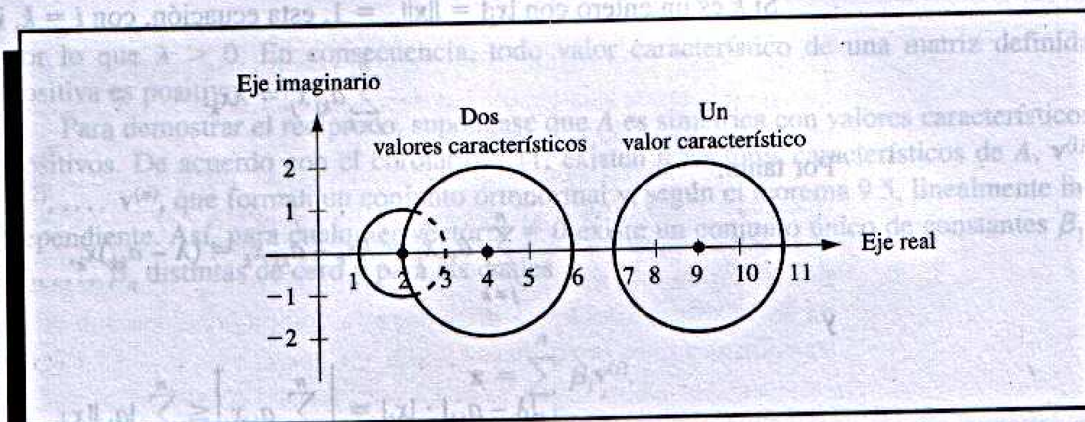
$$R_1 = \{z \in \mathcal{C} \mid |z - 4| \leq 2\},$$
$$R_2 = \{z \in \mathcal{C} \mid |z - 2| \leq 1\}$$

y

$$R_3 = \{z \in \mathcal{C} \mid |z - 9| \leq 2\}.$$

Puesto que R_1 y R_2 son ajenos a R_3, debe haber exactamente dos valores característicos dentro de $R_1 \cup R_2$ y uno dentro de R_3. Además, como $\rho(A) = \max_{1\le i\le 3} |\lambda_i|$, tenemos $7 \le \rho(A) \le 11$. ■

Figura 9.1



CONJUNTO DE EJERCICIOS 9.1

1. Obtenga los valores característicos y los vectores característicos asociados para las siguientes matrices de 3 × 3. ¿Hay un conjunto de tres vectores característicos linealmente independientes?

a. $A = \begin{bmatrix} 2 & -3 & 6 \\ 0 & 3 & -4 \\ 0 & 2 & -3 \end{bmatrix}$ **b.** $A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 1 \\ -1 & -1 & 2 \end{bmatrix}$

c. $A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}$ **d.** $A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$

e. $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ **f.** $A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$

2. Las matrices del ejercicio 1(c), (d), (e) y (f) son simétricas.

a. ¿Son definidas positivas?

b. Considere las matrices definidas positivas del inciso (a). Construya una matriz ortogonal Q para la cual $Q^t AQ = D$ sea una matriz diagonal, por medio de los vectores característicos encontrados en el ejercicio 1.

3. Aplique el teorema del círculo de Gerschgorin para determinar las cotas de los valores característicos de las siguientes matrices.

a. $\begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 1 \\ -1 & -1 & 2 \end{bmatrix}$ **b.** $\begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{bmatrix}$

c. $\begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 3 \end{bmatrix}$ **d.** $\begin{bmatrix} 4.75 & 2.25 & -0.25 \\ 2.25 & 4.75 & 1.25 \\ -0.25 & 1.25 & 4.75 \end{bmatrix}$

e. $\begin{bmatrix} -4 & 0 & 1 & 3 \\ 0 & -4 & 2 & 1 \\ 1 & 2 & -2 & 0 \\ 3 & 1 & 0 & -4 \end{bmatrix}$ f. $\begin{bmatrix} 1 & 0 & -1 & 1 \\ 2 & 2 & -1 & 1 \\ 0 & 1 & 3 & -2 \\ 1 & 0 & 1 & 4 \end{bmatrix}$

4. Demuestre que cuatro vectores cualesquiera en $\mathbb{R}^3$ son linealmente dependientes.
5. Sea $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ un conjunto de k vectores ortogonales diferentes de cero, demuestre que es un conjunto linealmente independiente.
6. Sea Q una matriz ortogonal.
   a. Demuestre que las columnas de Q forman un conjunto ortogonal de vectores.
   b. Demuestre que $\|Q\|_2 = 1$ y $\|Q^t\|_2 = 1$.
7. Sea $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ un conjunto de vectores ortonormales distintos de cero en $\mathbb{R}^n$ y $\mathbf{x} \in \mathbb{R}^n$. Determine los valores de c_k para $k = 1, 2, \ldots, n$, si
$$\mathbf{x} = \sum_{k=1}^{n} c_k \mathbf{v}_k.$$
8. Demuestre que si A es una matriz de $n \times n$ con n valores característicos distintos, entonces A tiene n vectores característicos linealmente independientes.
9. En el ejercicio 25 de la sección 6.6, se usó una matriz simétrica
$$A = \begin{bmatrix} 1.59 & 1.69 & 2.13 \\ 1.69 & 1.31 & 1.72 \\ 2.13 & 1.72 & 1.85 \end{bmatrix}$$
para describir las longitudes promedio del ala de las moscas de frutas que habían nacido del apareamiento de tres mutantes de moscas. El elemento a_{ij} representa la longitud promedio del ala de una mosca que nació de una mosca macho del tipo i y una mosca hembra del tipo j.
   a. Calcule los valores característicos y los vectores característicos asociados de esta matriz.
   b. Use el teorema 9.12 para contestar la pregunta planteada en el inciso (b) del ejercicio 25, sección 6.6, es decir, ¿es definida positiva esta matriz?
10. Una **matriz persimétrica** es aquella que es simétrica alrededor de ambas diagonales; es decir, una matriz A de $N \times N$, $A = (a_{ij})$ es persimétrica si $a_{ij} = a_{ji} = a_{N+1-i,\,N+1-j}$ para toda $i = 1, 2, \ldots, N$ y $j = 1, 2, \ldots, N$. Diversos problemas de la teoría de la comunicación tienen una solución que incluye los valores y vectores característicos de matrices de forma persimétrica. Por ejemplo, el vector característico correspondiente al valor característico mínimo de la matriz persimétrica de 4×4
$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$
da la respuesta al impulso unitario del canal de energía para una sucesión de errores de longitud 2, y después el peso mínimo de cualquier sucesión de error posible.
   a. Aplique el teorema del círculo de Gerschgorin para demostrar que, si A es la matriz dada arriba y si λ es un valor característico mínimo, entonces $|\lambda - 4| = \rho(A - 4I)$, donde ρ denota el radio espectral.
   b. Calcule el valor característico mínimo de la matriz A, obteniendo todos los valores característicos de $A - 4I$ y calculando su radio espectral. Después, calcule el vector característico correspondiente.
   c. Aplique el teorema del círculo de Gerschgorin para demostrar que si λ es el valor característico mínimo de la matriz

$$B = \begin{bmatrix} 3 & -1 & -1 & 1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 1 & -1 & -1 & 3 \end{bmatrix},$$

entonces $|\lambda - 6| = \rho(B - 6I)$.

d. Repita el inciso (b) usando la matriz B y el resultado del inciso (c).

9.2 Método de la potencia

El **método de la potencia** es una técnica iterativa que permite determinar el valor característico dominante de una matriz, es decir, el valor característico con mayor magnitud. Una ligera modificación del método permite determinar también otros valores característicos. Un aspecto útil del método de la potencia es que no sólo produce un valor característico, sino un vector característico asociado. De hecho, es frecuente que el método de la potencia se aplique para hallar un vector característico de un valor característico determinado por otros medios.

Para aplicar el método de la potencia supondremos que la matriz A de $n \times n$ tiene n valores característicos $\lambda_1, \lambda_2, \ldots, \lambda_n$ con un conjunto asociado de vectores característicos linealmente independientes $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(n)}\}$. Más aún, supondremos que A tiene exactamente un valor característico, λ_1, cuya magnitud es la mayor, por lo que $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \geq 0$.

Si $\mathbf{x}$ es un vector cualquiera en $\mathbb{R}^n$, el hecho de que $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \ldots, \mathbf{v}^{(n)}\}$ sea linealmente independiente implica que existen las constantes $\beta_1, \beta_2, \ldots, \beta_n$ con

$$\mathbf{x} = \sum_{j=1}^{n} \beta_j \mathbf{v}^{(j)}.$$

Al multiplicar ambos lados de esta ecuación por $A, A^2, \ldots, A^k$ obtenemos:

$$A\mathbf{x} = \sum_{j=1}^{n} \beta_j A\mathbf{v}^{(j)} = \sum_{j=1}^{n} \beta_j \lambda_j \mathbf{v}^{(j)},$$

$$A^2\mathbf{x} = \sum_{j=1}^{n} \beta_j \lambda_j A\mathbf{v}^{(j)} = \sum_{j=1}^{n} \beta_j \lambda_j^2 \mathbf{v}^{(j)},$$

$$\vdots$$

$$A^k\mathbf{x} = \sum_{j=1}^{n} \beta_j \lambda_j^k \mathbf{v}^{(j)}.$$

Si factorizamos λ_1^k en cada término de la derecha de la última ecuación, entonces

$$A^k\mathbf{x} = \lambda_1^k \sum_{j=1}^{n} \beta_j \left(\frac{\lambda_j}{\lambda_1}\right)^k \mathbf{v}^{(j)}.$$

Como $|\lambda_1| > |\lambda_j|$, para cualquier $j = 2, 3, \ldots, n$, tenemos $\lim_{k\to\infty} (\lambda_j/\lambda_1)^k = 0$, y

$$\lim_{k\to\infty} A^k\mathbf{x} = \lim_{k\to\infty} \lambda_1^k \beta_1 \mathbf{v}^{(1)}. \qquad (9.1)$$

Esta sucesión converge a cero si $|\lambda_1| < 1$ y diverge si $|\lambda_1| > 1$, naturalmente a condición de que $\beta_1 \neq 0$.

De la relación expresada en la ecuación (9.1) puede obtenerse una ventaja al escalar en forma adecuada las potencias de $A^k\mathbf{x}$ para asegurarnos de que el límite de la ecuación (9.1) sea finito y no cero. El escalamiento comienza seleccionando $\mathbf{x}$ como un vector unitario $\mathbf{x}^{(0)}$ en relación con $\|\cdot\|_\infty$ y una componente $x_{p_0}^{(0)}$ de $\mathbf{x}^{(0)}$ con

$$x_{p_0}^{(0)} = 1 = \|\mathbf{x}^{(0)}\|_\infty.$$

Sea $\mathbf{y}^{(1)} = A\mathbf{x}^{(0)}$ y definamos $\mu^{(1)} = y_{p_0}^{(1)}$. Con esta notación,

$$\mu^{(1)} = y_{p_0}^{(1)} = \frac{y_{p_0}^{(1)}}{x_{p_0}^{(0)}} = \frac{\beta_1\lambda_1 v_{p_0}^{(1)} + \sum_{j=2}^n \beta_j \lambda_j v_{p_0}^{(j)}}{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^n \beta_j v_{p_0}^{(j)}} = \lambda_1\left[\frac{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^n \beta_j(\lambda_j/\lambda_1)v_{p_0}^{(j)}}{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^n \beta_j v_{p_0}^{(j)}}\right].$$

Entonces, sea p_1 el menor entero tal que

$$|y_{p_1}^{(1)}| = \|\mathbf{y}^{(1)}\|_\infty,$$

y definamos $\mathbf{x}^{(1)}$ por medio de

$$\mathbf{x}^{(1)} = \frac{1}{y_{p_1}^{(1)}}\mathbf{y}^{(1)} = \frac{1}{y_{p_1}^{(1)}}A\mathbf{x}^{(0)}.$$

Entonces,

$$x_{p_1}^{(1)} = 1 = \|\mathbf{x}^{(1)}\|_\infty.$$

A continuación definimos

$$\mathbf{y}^{(2)} = A\mathbf{x}^{(1)} = \frac{1}{y_{p_1}^{(1)}}A^2\mathbf{x}^{(0)}$$

y

$$\mu^{(2)} = y_{p_1}^{(2)} = \frac{y_{p_1}^{(2)}}{x_{p_1}^{(1)}} = \frac{\left[\beta_1\lambda_1^2 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j\lambda_j^2 v_{p_1}^{(j)}\right] \Big/ y_{p_1}^{(1)}}{\left[\beta_1\lambda_1 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j\lambda_j v_{p_1}^{(j)}\right] \Big/ y_{p_1}^{(1)}}$$

$$= \lambda_1\left[\frac{\beta_1 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j(\lambda_j/\lambda_1)^2 v_{p_1}^{(j)}}{\beta_1 v_{p_1}^{(1)} + \sum_{j=2}^n \beta_j(\lambda_j/\lambda_1)v_{p_1}^{(j)}}\right].$$

Sea p_2 el entero más pequeño con

$$|y_{p_2}^{(2)}| = \|\mathbf{y}^{(2)}\|_\infty.$$

y definamos

$$\mathbf{x}^{(2)} = \frac{1}{y_{p_2}^{(2)}}\mathbf{y}^{(2)} = \frac{1}{y_{p_2}^{(2)}}A\mathbf{x}^{(1)} = \frac{1}{y_{p_2}^{(2)}y_{p_1}^{(1)}}A^2\mathbf{x}^{(0)}.$$

De modo semejante definimos las sucesiones de los vectores $\{\mathbf{x}^{(m)}\}_{m=0}^{\infty}$ y $\{\mathbf{y}^{(m)}\}_{m=1}^{\infty}$, e inductivamente hacemos lo mismo con una sucesión de escalares $\{\mu^{(m)}\}_{m=1}^{\infty}$ mediante

$$\mathbf{y}^{(m)} = A\mathbf{x}^{(m-1)},$$

$$\mu^{(m)} = y_{p_{m-1}}^{(m)} = \lambda_1\left[\frac{\beta_1 v_{p_{m-1}}^{(1)} + \sum_{j=2}^{n}(\lambda_j/\lambda_1)^m\beta_j v_{p_{m-1}}^{(j)}}{\beta_1 v_{p_{m-1}}^{(1)} + \sum_{j=2}^{n}(\lambda_j/\lambda_1)^{m-1}\beta_j v_{p_{m-1}}^{(j)}}\right] \tag{9.2}$$

y

$$\mathbf{x}^{(m)} = \frac{\mathbf{y}^{(m)}}{y_{p_m}^{(m)}} = \frac{A^m\mathbf{x}^{(0)}}{\prod_{k=1}^{m} y_{p_k}^{(k)}},$$

donde cada paso p_m sirve para representar el entero más pequeño para el cual

$$|y_{p_m}^{(m)}| = \|\mathbf{y}^{(m)}\|_\infty.$$

Al examinar la ecuación (9.2) vemos que, como $|\lambda_j/\lambda_1| < 1$ para cualquier $j = 2, 3, \ldots, n$, $\lim_{m\to\infty}\mu^{(m)} = \lambda_1$, siempre y cuando elijamos $\mathbf{x}^{(0)}$ de modo que $\beta_1 \neq 0$. Más aún, la sucesión de vectores $\{\mathbf{x}^{(m)}\}_{m=0}^{\infty}$ converge al vector característico asociado a λ_1, que tiene norma uno l_∞.

El método de la potencia tiene la desventaja de que al inicio no se sabe si la matriz tiene o no un solo valor característico dominante. Tampoco se sabe cómo seleccionar $\mathbf{x}^{(0)}$ para estar seguros de que su representación mediante vectores característicos de la matriz contenga una contribución distinta de cero del vector característico asociado al valor característico dominante, en caso de que exista.

En el algoritmo 9.1 se ejecuta el método de la potencia.

ALGORITMO 9.1

Método de la potencia

Para aproximar el valor característico dominante y el vector característico asociado de la matriz A de $n \times n$ con un vector $\mathbf{x}$ distinto de cero:

ENTRADA dimensión n; matriz A; vector $\mathbf{x}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA valor característico aproximado μ; vector característico aproximado $\mathbf{x}$ (con $\|\mathbf{x}\|_\infty = 1$) o bien un mensaje de que se rebasó el número máximo de iteraciones.

Paso 1 Tome $k = 1$.

Paso 2 Obtenga el entero p más pequeño con $1 \leq p \leq n$ y $|x_p| = \|\mathbf{x}\|_\infty$.

Paso 3 Tome $\mathbf{x} = \mathbf{x}/x_p$.

Paso 4 Mientras ($k \leq N$) haga los pasos 5-11.

Paso 5 Tome $\mathbf{y} = A\mathbf{x}$.

Paso 6 Tome $\mu = y_p$.

Paso 7 Obtenga el entero p más pequeño con $1 \leq p \leq n$ y $|y_p| = \|\mathbf{y}\|_\infty$.

Paso 8 Si $y_p = 0$, entonces SALIDA ('Vector característico', $\mathbf{x}$);
SALIDA ('A tiene el valor característico 0, seleccione un nuevo vector $\mathbf{x}$ y reinicie');
PARAR.

Paso 9 Tome $ERR = \|\mathbf{x} - (\mathbf{y}/y_p)\|_\infty$;
$\mathbf{x} = \mathbf{y}/y_p$.

Paso 10 Si $ERR < TOL$, entonces SALIDA (μ, $\mathbf{x}$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 11 Tome $k = k + 1$.

Paso 12 SALIDA ('Excedido el número máximo de iteraciones');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Al seleccionar, en el paso 7, el entero más pequeño p_m para el cual $|y_{p_m}^{(m)}| = \|\mathbf{y}^{(m)}\|_\infty$ generalmente se garantiza que este índice finalmente se vuelva invariable. La rapidez con que $\{\mu^{(m)}\}_{m=1}^{\infty}$ converge a λ_1 se determina mediante las razones $|\lambda_j/\lambda_1|^m$, para $j = 2, 3, \ldots, n$, y, particularmente, mediante $|\lambda_2/\lambda_1|^m$. La rapidez de convergencia es $O(|\lambda_2/\lambda_1|^m)$ (véase [IK, p. 148]), por lo cual hay una constante k tal que para m grande,

$$|\mu^{(m)} - \lambda_1| \approx k \left|\frac{\lambda_2}{\lambda_1}\right|^m,$$

lo anterior significa que

$$\lim_{m\to\infty} \frac{|\mu^{(m+1)} - \lambda_1|}{|\mu^{(m)} - \lambda_1|} \approx \left|\frac{\lambda_2}{\lambda_1}\right| < 1.$$

Entonces, la sucesión $\{\mu^{(m)}\}$ converge linealmente a λ_1, y, por tanto, el procedimiento de Aitken Δ^2 que se explicó en la sección 2.5 puede servir para agilizar la convergencia. La implantación del procedimiento Δ^2 en el algoritmo 9.1 se logra modificándolo como sigue:

Paso 1 Tome $k = 1$;
$\mu_0 = 0$;
$\mu_1 = 0$.

Paso 6 Tome $\mu = y_p$;

$$\hat{\mu} = \mu_0 - \frac{(\mu_1 - \mu_0)^2}{\mu - 2\mu_1 + \mu_0}.$$

Paso 10 Si $ERR < TOL$ y $k \geq 4$, entonces SALIDA $(\hat{\mu}, \mathbf{x})$;
PARAR.

Paso 11 Tome $k = k + 1$;
$\mu_0 = \mu_1$;
$\mu_1 = \mu$.

No es necesario que la matriz tenga valores característicos distintos para que converja el método de la potencia. Si el valor característico dominante y único, λ_1, tiene una multiplicidad r mayor que 1, y si $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(r)}$ son vectores característicos linealmente independientes asociados a λ_1, el procedimiento todavía convergirá en λ_1. En este caso, la sucesión de vectores $\{\mathbf{x}^{(m)}\}_{m=0}^{\infty}$ convergirá a un vector característico de λ_1 de l_∞, con norma uno, que es una combinación lineal de $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(r)}$ y depende de la elección del vector inicial $\mathbf{x}^{(0)}$.

EJEMPLO 1 La matriz

$$A = \begin{bmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{bmatrix}$$

tiene los valores característicos $\lambda_1 = 6$, $\lambda_2 = 3$ y $\lambda_3 = 2$. En consecuencia, el método de la potencia descrito en el algoritmo 9.1 convergirá. Sea $\mathbf{x}^{(0)} = (1, 1, 1)^t$, entonces

$$\mathbf{y}^{(1)} = A\mathbf{x}^{(0)} = (10, 8, 1)^t,$$

por lo que

$$\|\mathbf{y}^{(1)}\|_\infty = 10, \quad \mu^{(1)} = y_1^{(1)} = 10 \quad \text{y} \quad \mathbf{x}^{(1)} = \frac{\mathbf{y}^{(1)}}{10} = (1, 0.8, 0.1)^t.$$

Continuando en esta forma se generan los valores de la tabla 9.1, donde $\hat{\mu}^{(m)}$ representa la secuencia generada por el procedimiento de Aitken Δ^2. En esta etapa, una aproximación al valor característico dominante, 6, es $\hat{\mu}^{(10)} = 6.000000$ con el vector característico unitario aproximado $(1, 0.714316, -0.249895)^t$. Aunque la aproximación al valor caracte-

Tabla 9.1

m	$(\mathbf{x}^{(m)})^t$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$
0	(1, 1, 1)		
1	(1, 0.8, 0.1)	10	6.266667
2	(1, 0.75, −0.111)	7.2	6.062473
3	(1, 0.730769, −0.188803)	6.5	6.015054
4	(1, 0.722200, −0.220850)	6.230769	6.004202
5	(1, 0.718182, −0.235915)	6.111000	6.000855
6	(1, 0.716216, −0.243095)	6.054546	6.000240
7	(1, 0.715247, −0.246588)	6.027027	6.000058
8	(1, 0.714765, −0.248306)	6.013453	6.000017
9	(1, 0.714525, −0.249157)	6.006711	6.000003
10	(1, 0.714405, −0.249579)	6.003352	6.000000
11	(1, 0.714346, −0.249790)	6.001675	
12	(1, 0.714316, −0.249895)	6.000837	

rístico es correcta hasta las cifras enumeradas, la aproximación al vector característico es mucho menos cercana al verdadero vector característico $(1, 0.714286, -0.25)^t$.

■

Cuando A es simétrica, podemos hacer una variación en la elección de los vectores $\mathbf{x}^{(m)}$, $\mathbf{y}^{(m)}$ y escalares $\mu^{(m)}$ para mejorar significativamente la razón de convergencia entre la sucesión $\{\mu^{(m)}\}_{m=1}^{\infty}$ y el valor característico dominante λ_1. De hecho, aunque la razón de convergencia del método general de potencia es $O(|\lambda_2/\lambda_1|^m)$, la del método modificado que se da en el algoritmo 9.2 para matrices simétricas es $O(|\lambda_2/\lambda_1|^{2m})$. (Véase [IK, pp. 149 ss.].) La sucesión $\{\mu^{(m)}\}$ es todavía linealmente convergente, de modo que podemos volver a aplicar el procedimiento Δ^2 de Aitken.

ALGORITMO 9.2

Método de la potencia simétrica

Para aproximar el valor característico dominante y el vector característico asociado de una matriz simétrica A de $n \times n$, dado un vector $\mathbf{x}$ diferente de cero:

ENTRADA dimensión n; matriz A; vector $\mathbf{x}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA valor característico aproximado μ; vector característico $\mathbf{x}$ aproximado (con $\|\mathbf{x}\|_2 = 1$) o bien un mensaje de que se rebasó el número máximo de iteraciones.

Paso 1 Tome $k = 1$;
$\mathbf{x} = \mathbf{x}/\|\mathbf{x}\|_2$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3-8.

Paso 3 Tome $\mathbf{y} = A\mathbf{x}$.

Paso 4 Tome $\mu = \mathbf{x}^t\mathbf{y}$.

Paso 5 Si $\|\mathbf{y}\|_2 = 0$, entonces SALIDA ('Vector característico', $\mathbf{x}$);
SALIDA ('A tiene el valor característico 0, seleccione un nuevo vector $\mathbf{x}$ y reinicie');
PARAR.

Paso 6 Tome $ERR = \left\| \mathbf{x} - \dfrac{\mathbf{y}}{\|\mathbf{y}\|_2} \right\|_2$;
$\mathbf{x} = \mathbf{y}/\|\mathbf{y}\|_2$.

Paso 7 Si $ERR < TOL$, entonces SALIDA (μ, $\mathbf{x}$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 8 Tome $k = k + 1$.

Paso 9 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR.

■

EJEMPLO 2 La matriz

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{bmatrix}$$

es simétrica con los valores característicos $\lambda_1 = 6$, $\lambda_2 = 3$ y $\lambda_3 = 1$. La tabla 9.2 enumera los resultados del método de la potencia, y los resultados de la tabla 9.3 provienen del método de la potencia simétrica, suponiendo en cada caso que $\mathbf{y}^{(0)} = \mathbf{x}^{(0)} = (1, 0, 0)^t$. Observe la mejora significativa proporcionada por el método de la potencia simétrica. Las aproximaciones a los vectores característicos producidos con el método de la potencia convergen en $(1, -1, 1)^t$, un vector con $\|(1, -1, 1)^t\|_\infty = 1$. En el método de la potencia simétrica, la convergencia es hacia el vector paralelo $(\sqrt{3}/3, -\sqrt{3}/3, \sqrt{3}/3)^t$, con $\|(\sqrt{3}/3, -\sqrt{3}/3, \sqrt{3}/3)^t\|_2 = 1$. ■

Tabla 9.2

m	$(\mathbf{y}^{(m)})^t$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$	$(\mathbf{x}^{(m)})^t$ con $\|\mathbf{x}^{(m)}\|_\infty = 1$
0				$(1, 0, 0)$
1	$(4, -1, 1)$	4		$(1, -0.25, 0.25)$
2	$(4.5, -2.25, 2.25)$	4.5	7	$(1, -0.5, 0.5)$
3	$(5, -3.5, 3.5)$	5	6.2	$(1, -0.7, 0.7)$
4	$(5.4, -4.5, 4.5)$	5.4	6.047617	$(1, -8.33\overline{3}, 0.833\overline{3})$
5	$(5.66\overline{6}, -5.166\overline{6}, 5.166\overline{6})$	$5.66\overline{6}$	6.011767	$(1, -0.911765, 0.911765)$
6	$(5.823529, -5.558824, 5.558824)$	5.823529	6.002931	$(1, -0.954545, 0.954545)$
7	$(5.909091, -5.772727, 5.772727)$	5.909091	6.000733	$(1, -0.976923, 0.976923)$
8	$(5.953846, -5.884615, 5.884615)$	5.953846	6.000184	$(1, -0.988372, 0.988372)$
9	$(5.976744, -5.941861, 5.941861)$	5.976744		$(1, -0.994163, 0.994163)$
10	$(5.988327, -5.970817, 5.970817)$	5.988327		$(1, -0.997076, 0.997076)$

Tabla 9.3

m	$(\mathbf{y}^{(m)})^t$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$	$(\mathbf{x}^{(m)})^t$ con $\|\mathbf{x}^{(m)}\|_2 = 1$
0	$(1, 0, 0)$			$(1, 0, 0)$
1	$(4, -1, 1)$	4	7	$(0.942809, -0.235702, 0.235702)$
2	$(4.242641, -2.121320, 2.121320)$	5	6.047619	$(0.816497, -0.408248, 0.408248)$
3	$(4.082483, -2.857738, 2.857738)$	5.666667	6.002932	$(0.710669, -0.497468, 0.497468)$
4	$(3.837613, -3.198011, 3.198011)$	5.909091	6.000183	$(0.646997, -0.539164, 0.539164)$
5	$(3.666314, -3.342816, 3.342816)$	5.976744	6.000012	$(0.612836, -0.558763, 0.558763)$
6	$(3.568871, -3.406650, 3.406650)$	5.994152	6.000000	$(0.595247, -0.568190, 0.568190)$
7	$(3.517370, -3.436200, 3.436200)$	5.998536	6.000000	$(0.586336, -0.572805, 0.572805)$
8	$(3.490952, -3.450359, 3.450359)$	5.999634		$(0.581852, -0.575086, 0.575086)$
9	$(3.477580, -3.457283, 3.457283)$	5.999908		$(0.579603, -0.576220, 0.576220)$
10	$(3.470854, -3.460706, 3.460706)$	5.999977		$(0.578477, -0.576786, 0.576786)$

El siguiente teorema ofrece una cota de error para aproximar los valores característicos de una matriz simétrica.

Teorema 9.14 Si A es una matriz simétrica de $n \times n$ con los valores característicos $\lambda_1, \lambda_2, \ldots, \lambda_n$ y $\|A\mathbf{x} - \lambda\mathbf{x}\|_2 < \varepsilon$ para algún vector $\mathbf{x}$ con $\|\mathbf{x}\|_2 = 1$ y con el número real λ, entonces

$$\min_{1 \le j \le n} |\lambda_j - \lambda| < \varepsilon. \quad \blacksquare$$

Demostración Supongamos que $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$ forman un conjunto ortonormal de vectores característicos de A asociados, respectivamente, con los valores característicos $\lambda_1, \lambda_2, \ldots, \lambda_n$. De acuerdo con los teoremas 9.5 y 9.2, para algún conjunto único de constantes $\beta_1, \beta_2, \ldots, \beta_n$, podemos expresar $\mathbf{x}$ como

$$\mathbf{x} = \sum_{j=1}^{n} \beta_j \mathbf{v}^{(j)}.$$

Por tanto,

$$\|A\mathbf{x} - \lambda\mathbf{x}\|_2^2 = \left\| \sum_{j=1}^{n} \beta_j(\lambda_j - \lambda)\mathbf{v}^{(j)} \right\|_2^2 = \sum_{j=1}^{n} |\beta_j|^2 |\lambda_j - \lambda|^2 \ge \min_{1 \le j \le n} |\lambda_j - \lambda|^2 \sum_{j=1}^{n} |\beta_j|^2.$$

Pero

$$\sum_{j=1}^{n} |\beta_j|^2 = \|\mathbf{x}\|_2^2 = 1,$$

por lo que

$$\varepsilon \ge \|A\mathbf{x} - \lambda\mathbf{x}\|_2 > \min_{1 \le j \le n} |\lambda_j - \lambda|. \quad \blacksquare\ \blacksquare\ \blacksquare$$

El **método de la potencia inversa** es una modificación del método de la potencia que ofrece una convergencia más rápida. Se usa para determinar el valor característico de A más cercano a un número q específico.

Supongamos que la matriz A tiene los valores característicos $\lambda_1, \ldots, \lambda_n$ con los vectores característicos linealmente independientes $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$. Consideremos la matriz $(A - qI)^{-1}$, donde $q \ne \lambda_i$ para $i = 1, 2, \ldots, n$. Los valores característicos de $(A - qI)^{-1}$ son

$$\frac{1}{\lambda_1 - q}, \quad \frac{1}{\lambda_2 - q}, \quad \ldots, \quad \frac{1}{\lambda_n - q},$$

con los vectores característicos $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$ (véase el ejercicio 9 de la sección 7.2). Al aplicar el método de la potencia en $(A - qI)^{-1}$ obtenemos

$$\mathbf{y}^{(m)} = (A - qI)^{-1}\mathbf{x}^{(m-1)},$$

$$\mu^{(m)} = y_{p_{m-1}}^{(m)} = \frac{y_{p_{m-1}}^{(m)}}{y_{p_{m-1}}^{(m-1)}} = \frac{\sum_{j=1}^{n} \beta_j \dfrac{1}{(\lambda_j - q)^m} v_{p_{m-1}}^{(j)}}{\sum_{j=1}^{n} \beta_j \dfrac{1}{(\lambda_j - q)^{m-1}} v_{p_{m-1}}^{(j)}} \tag{9.3}$$

y

$$\mathbf{x}^{(m)} = \frac{\mathbf{y}^{(m)}}{y_{p_m}^{(m)}},$$

donde p_m representa en cada paso el entero más pequeño para el cual $|y_{p_m}^{(m)}| = \|y^{(m)}\|_\infty$. La sucesión $\{\mu^{(m)}\}$ de la ecuación (9.3) converge a $1/(\lambda_k - q)$, donde

$$\frac{1}{|\lambda_k - q|} = \max_{1 \le i \le n} \frac{1}{|\lambda_i - q|}$$

y $\lambda_k \approx q + 1/\mu^{(m)}$ es el valor característico de A más cercano a q.

Cuando k se conoce, la ecuación (9.3) puede escribirse así

$$\mu^{(m)} = \frac{1}{\lambda_k - q} \left[\frac{\beta_k v_{p_{m-1}}^{(k)} + \sum_{\substack{j=1 \\ j \ne k}}^{n} \beta_j \left[\frac{\lambda_k - q}{\lambda_j - q}\right]^m v_{p_{m-1}}^{(j)}}{\beta_k v_{p_{m-1}}^{(k)} + \sum_{\substack{j=1 \\ j \ne k}}^{n} \beta_j \left[\frac{\lambda_k - q}{\lambda_j - q}\right]^{m-1} v_{p_{m-1}}^{(j)}} \right]. \tag{9.4}$$

Por tanto, la elección de q determina la convergencia siempre y cuando $1/(\lambda_k - q)$ sea un valor característico dominante y único de $(A - qI)^{-1}$ (aunque puede ser un valor característico múltiple). Cuanto más se acerque q a un valor característico λ_k de A, más rápida será la convergencia, porque ésta es del orden de

$$O\left(\left|\frac{(\lambda - q)^{-1}}{(\lambda_k - q)^{-1}}\right|^m\right) = O\left(\left|\frac{(\lambda_k - q)}{(\lambda - q)}\right|^m\right),$$

donde λ representa el valor característico de A, que es el segundo más cercano a q.

El cálculo del vector $\mathbf{y}^{(m)}$ se obtiene de la ecuación

$$(A - qI)\mathbf{y}^{(m)} = \mathbf{x}^{(m-1)}.$$

En términos generales, para resolver este sistema se utiliza la eliminación gaussiana con pivoteo.

Aunque el método de la potencia inversa requiere la solución de un sistema de $n \times n$ en cada paso, los múltiplos pueden guardarse para reducir los cálculos. La selección de q puede tener como base el teorema del círculo de Gerschgorin o cualquier otro medio para localizar el valor característico.

El algoritmo 9.3 calcula q a partir de una aproximación inicial $\mathbf{x}^{(0)}$ de un vector característico por medio de

$$q = \frac{\mathbf{x}^{(0)t} A \mathbf{x}^{(0)}}{\mathbf{x}^{(0)t} \mathbf{x}^{(0)}}.$$

Esta elección de q proviene de la observación de que, si $\mathbf{x}$ es un vector característico de A respecto al valor característico λ, entonces $A\mathbf{x} = \lambda\mathbf{x}$. Por tanto, $\mathbf{x}^t A\mathbf{x} = \lambda \mathbf{x}^t \mathbf{x}$ y

$$\lambda = \frac{\mathbf{x}^t A \mathbf{x}}{\mathbf{x}^t \mathbf{x}} = \frac{\mathbf{x}^t A \mathbf{x}}{\|\mathbf{x}\|_2^2}.$$

Si q está cerca de un valor característico, la convergencia será muy rápida, pero en el paso 6 deberá emplearse una técnica de pivoteo para evitar la contaminación por el error de redondeo.

A menudo, el algoritmo 9.3 se usa para aproximar un vector característico cuando se conoce un valor característico q aproximado.

ALGORITMO 9.3

Método de la potencia inversa

Para aproximar un valor característico y un vector característico asociado de la matriz A de $n \times n$, dado un vector $\mathbf{x}$ distinto de cero:

ENTRADA dimensión n; matriz A; vector $\mathbf{x}$; tolerancia TOL; número máximo de iteraciones N.

SALIDA valor característico aproximado μ; vector característico aproximado $\mathbf{x}$ (con $\|\mathbf{x}\|_\infty = 1$) o un mensaje de que se rebasó el número máximo de iteraciones.

Paso 1 Tome $q = \dfrac{\mathbf{x}^t A\mathbf{x}}{\mathbf{x}^t\mathbf{x}}$.

Paso 2 Tome $k = 1$.

Paso 3 Obtenga el entero p más pequeño con $1 \le p \le n$ y $|x_p| = \|\mathbf{x}\|_\infty$.

Paso 4 Tome $\mathbf{x} = \mathbf{x}/x_p$.

Paso 5 Mientras ($k \le N$) haga los pasos 6-12.

Paso 6 Resuelva el sistema lineal $(A - qI)\mathbf{y} = \mathbf{x}$.

Paso 7 Si el sistema no tiene una solución única, entonces
SALIDA ('q es un valor característico', q);
PARAR.

Paso 8 Tome $\mu = y_p$.

Paso 9 Obtenga el entero p más pequeño con $1 \le p \le n$ y $|y_p| = \|\mathbf{y}\|_\infty$.

Paso 10 Tome $ERR = \|\mathbf{x} - (\mathbf{y}/y_p)\|_\infty$;
$\mathbf{x} = \mathbf{y}/y_p$.

Paso 11 Si $ERR < TOL$, entonces tome $\mu = (1/\mu) + q$;
SALIDA (μ, $\mathbf{x}$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 12 Tome $k = k + 1$.

Paso 13 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ∎

Como la convergencia del método de potencia inversa es lineal, podemos usar nuevamente el procedimiento Δ^2 de Aitken para acelerar la convergencia. El siguiente ejemplo ilustra la convergencia rápida del método de la potencia inversa si q está cerca de un valor característico.

EJEMPLO 3 La matriz

$$A = \begin{bmatrix} -4 & 14 & 0 \\ -5 & 13 & 0 \\ -1 & 0 & 2 \end{bmatrix}$$

se consideró en el ejemplo 1. El algoritmo 9.1 dio la aproximación $\mu^{(12)} = 6.000837$ por medio de $\mathbf{x}^{(0)} = (1, 1, 1)^t$. Con $\mathbf{x}^{(0)} = (1, 1, 1)^t$, tenemos

$$q = \frac{\mathbf{x}^{(0)t} A\mathbf{x}^{(0)}}{\mathbf{x}^{(0)t}\mathbf{x}^{(0)}} = \frac{19}{3} = 6.333333.$$

En la tabla 9.4 se incluyen los resultados obtenidos al aplicar el algoritmo 9.3 y el método Δ^2 de Aitken a $\mu^{(m)}$. ■

Tabla 9.4

m	$\mathbf{x}^{(m)t}$	$\mu^{(m)}$	$\hat{\mu}^{(m)}$
0	(1, 1, 1)		
1	(1, 0.720727, −0.194042)	6.183183	6.000116
2	(1, 0.715518, −0.245052)	6.017244	6.000004
3	(1, 0.714409, −0.249522)	6.001719	6.000004
4	(1, 0.714298, −0.249953)	6.000175	6.000003
5	(1, 0.714287, −0.250000)	6.000021	
6	(1, 0.714286, −0.249999)	6.000005	

Si A es simétrica, entonces para todo número real q, $(A - qI)^{-1}$ también es simétrica y, por tanto, también podemos aplicar el método de la potencia simétrica, es decir, el algoritmo 9.2, para acelerar la convergencia de $(A - qI)^{-1}$ en

$$O\left(\left|\frac{(\lambda_k - q)}{(\lambda - q)}\right|^{2m}\right).$$

Existen muchas técnicas para obtener aproximaciones a los otros valores característicos de una matriz, una vez calculada una aproximación al valor característico dominante. Limitaremos nuestra presentación a los **métodos de deflación**.

En un método de deflación debemos formar una matriz B nueva cuyos valores característicos sean iguales a los de A, salvo que el valor característico dominante de A se reemplaza con el valor característico 0 en B. El siguiente resultado justifica el procedimiento. La demostración de este teorema se encuentra en [Wil2, p. 596].

Teorema 9.15 Supongamos que $\lambda_1, \lambda_2, \ldots, \lambda_n$ son valores característicos de A con los vectores característicos asociados $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$ y que λ_1 tiene multiplicidad 1. Sea $\mathbf{x}$ un vector con $\mathbf{x}^t\mathbf{v}^{(1)} = 1$. Entonces, la matriz

$$B = A - \lambda_1\mathbf{v}^{(1)}\mathbf{x}^t$$

tiene los valores característicos $0, \lambda_2, \lambda_3, \ldots, \lambda_n$ con los vectores característicos asociados $\mathbf{v}^{(1)}, \mathbf{w}^{(2)}, \mathbf{w}^{(3)}, \ldots, \mathbf{w}^{(n)}$, donde $\mathbf{v}^{(i)}$ y $\mathbf{w}^{(i)}$ están relacionados por la ecuación

$$\mathbf{v}^{(i)} = (\lambda_i - \lambda_1)\mathbf{w}^{(i)} + \lambda_1(\mathbf{x}^t\mathbf{w}^{(i)})\,\mathbf{v}^{(1)}, \tag{9.5}$$

para cualquier $i = 2, 3, \ldots, n$. ■

Hay muchas elecciones del vector **x** que podríamos emplear en el teorema 9.15. La **deflación de Wielandt** se inicia con la definición

$$\mathbf{x} = \frac{1}{\lambda_1 v_i^{(1)}} (a_{i1}, a_{i2}, \ldots, a_{in})^t, \tag{9.6}$$

donde $v_i^{(1)}$ es una coordenada del vector característico $\mathbf{v}^{(1)}$ distinta de cero y los valores a_{i1}, $a_{i2}, \ldots, a_{in}$ son los elementos del i-ésimo renglón de A.

Con esta definición,

$$\mathbf{x}^t\mathbf{v}^{(1)} = \frac{1}{\lambda_1 v_i^{(1)}} [a_{i1}, a_{i2}, \ldots, a_{in}](v_1^{(1)}, v_2^{(1)}, \ldots, v_n^{(1)})^t = \frac{1}{\lambda_1 v_i^{(1)}} \sum_{j=1}^{n} a_{ij} v_j^{(1)},$$

donde la suma es la i-ésima coordenada del producto $A\mathbf{v}^{(1)}$. Dado que $A\mathbf{v}^{(1)} = \lambda_1\mathbf{v}^{(1)}$, esto significa que

$$\sum_{j=1}^{n} a_{ij} v_j^{(1)} = \lambda_1 v_i^{(1)},$$

lo que implica que

$$\mathbf{x}^t\mathbf{v}^{(1)} = \frac{1}{\lambda_1 v_i^{(1)}} (\lambda_1 v_i^{(1)}) = 1.$$

Por tanto, **x** satisface las hipótesis del teorema 9.15. Más aún (véase el ejercicio 12), el i-ésimo renglón de $B = A - \lambda_1\mathbf{v}^{(1)}\mathbf{x}^t$ consta por completo de elementos cero.

Si $\lambda \neq 0$ es un valor característico con vector característico asociado **w**, la relación $B\mathbf{w} = \lambda\mathbf{w}$ implica que la i-ésima coordenada de **w** también debe ser cero. En consecuencia, la columna de la matriz B no aporta nada al producto $B\mathbf{w} = \lambda\mathbf{w}$. Así, la matriz B puede ser reemplazada por una matriz B' $(n-1) \times (n-1)$ obtenida al suprimir en B el i-ésimo renglón y la i-ésima columna. La matriz B' tiene los valores característicos $\lambda_2, \lambda_3, \ldots, \lambda_n$. Si $|\lambda_2| > |\lambda_3|$, se aplica nuevamente el método de la potencia a la matriz B' para determinar este nuevo valor característico dominante y un vector característico $\mathbf{w}^{(2)'}$ asociado a λ_2, respecto a la matriz B'. Si quiere obtener el vector característico asociado $\mathbf{w}^{(2)}$ de la matriz B, introduzca una coordenada cero entre las coordenadas $w_{i-1}^{(2)'}$ y $w_i^{(2)'}$ del vector dimensional $(n-1)$ $\mathbf{w}^{(2)'}$ y luego calcule $\mathbf{v}^{(2)}$ mediante la ecuación (9.5).

EJEMPLO 4 Con base en el ejemplo 2, sabemos que la matriz

$$A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{bmatrix}$$

tiene los valores característicos $\lambda_1 = 6$, $\lambda_2 = 3$ y $\lambda_3 = 1$. Suponiendo que ya se calcularon el valor característico dominante $\lambda_1 = 6$ y el vector característico unitario asociado $\mathbf{v}^{(1)} = (1, -1, 1)^t$, el procedimiento que acabamos de describir para obtener λ_2 se realiza así:

$$\mathbf{x} = \frac{1}{6}\begin{bmatrix} 4 \\ -1 \\ 1 \end{bmatrix} = \left(\frac{2}{3}, -\frac{1}{6}, \frac{1}{6}\right)^t,$$

$$\mathbf{v}^{(1)}\mathbf{x}^t = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \left[\frac{2}{3}, -\frac{1}{6}, \frac{1}{6}\right] = \begin{bmatrix} \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \\ -\frac{2}{3} & \frac{1}{6} & -\frac{1}{6} \\ \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \end{bmatrix}$$

y

$$B = A - \lambda_1 \mathbf{v}^{(1)}\mathbf{x}^t = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{bmatrix} - 6 \begin{bmatrix} \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \\ -\frac{2}{3} & \frac{1}{6} & -\frac{1}{6} \\ \frac{2}{3} & -\frac{1}{6} & \frac{1}{6} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 3 & 2 & -1 \\ -3 & -1 & 2 \end{bmatrix}.$$

Al eliminar el primer renglón y la primera columna, obtenemos

$$B' = \begin{bmatrix} 2 & -1 \\ -1 & -2 \end{bmatrix},$$

que tiene los valores característicos $\lambda_2 = 3$ y $\lambda_3 = 1$. Para $\lambda_2 = 3$, el vector característico $\mathbf{w}^{(2)'}$ puede obtenerse resolviendo el sistema lineal de segundo orden

$$(B' - 3I)\mathbf{w}^{(2)'} = \mathbf{0},$$

que da por resultado

$$\mathbf{w}^{(2)'} = (1, -1)^t.$$

Al agregar un cero en el primer componente se tiene $\mathbf{w}^{(2)} = (0, 1, -1)^t$, y la ecuación (9.5) implica que tenemos el vector característico $\mathbf{v}^{(2)}$ de A correspondiente a $\lambda_2 = 3$:

$$\mathbf{v}^{(2)} = (3-6)(0, 1, -1)^t + 6\left[\left(\frac{2}{3}, -\frac{1}{6}, \frac{1}{6}\right)(0, 1, -1)^t\right](1, -1, 1)^t = (-2, -1, 1)^t. \quad \blacksquare$$

Aunque este proceso de deflación puede servir para obtener aproximaciones a todos los valores y vectores característicos de una matriz, el proceso es susceptible al error de redondeo. Si lo empleamos para calcular todos los valores característicos de una matriz, las aproximaciones conseguidas deberán usarse como valores iniciales del método de la potencia inversa aplicado a la matriz original. Esto garantiza que las aproximaciones converjan en los valores característicos de la matriz original, no a los de la matriz reducida, que probablemente contenga errores. Cuando se necesitan todos los valores característicos de una matriz, hay que usar las técnicas de la sección 9.4, basadas en transformaciones de similitud.

Terminaremos esta sección con el algoritmo 9.4, que calcula el segundo valor característico más dominante y el vector característico asociado de una matriz, una vez determinados el valor característico dominante y el vector característico dominante asociado.

ALGORITMO 9.4

Deflación de Wielandt

Para aproximar el segundo valor característico más dominante y el vector característico asociado de la matriz A de $n \times n$, dada una aproximación λ al valor característico dominante, se utiliza una aproximación $\mathbf{v}$ al vector característico correspondiente y un vector $\mathbf{x} \in \mathbb{R}^{n-1}$:

ENTRADA dimensión n; matriz A; valor característico aproximado λ con el vector característico $\mathbf{v} \in \mathbb{R}^n$; vector $\mathbf{x} \in \mathbb{R}^{n-1}$, tolerancia TOL; numero máximo de iteraciones N.

SALIDA valor característico aproximado μ; vector característico aproximado $\mathbf{u}$ o un mensaje de que el método falla.

Paso 1 Sea i el entero más pequeño con $1 \le i \le n$ y $|v_i| = \max_{1\le j\le n}|v_j|$.

Paso 2 Si $i \neq 1$, entonces
para $k = 1, \ldots, i-1$
para $j = 1, \ldots, i-1$

$$\text{tome } b_{kj} = a_{kj} - \frac{v_k}{v_i} a_{ij}.$$

Paso 3 Si $i \neq 1$ e $i \neq n$, entonces
para $k = i, \ldots, n-1$
para $j = 1, \ldots, i-1$

$$\text{tome } b_{kj} = a_{k+1,j} - \frac{v_{k+1}}{v_i} a_{ij};$$

$$b_{jk} = a_{j,k+1} - \frac{v_j}{v_i} a_{i,k+1}.$$

Paso 4 Si $i \neq n$, entonces
para $k = i, \ldots, n-1$
para $j = i, \ldots, n-1$

$$\text{tome } b_{kj} = a_{k+1,j+1} - \frac{v_{k+1}}{v_i} a_{i,j+1}.$$

Paso 5 Realice el método de la potencia en la matriz $(n-1) \times (n-1)$ $B' = (b_{kj})$ con $\mathbf{x}$ como aproximación inicial.

Paso 6 Si el método falla, entonces SALIDA ('Método falla');
PARAR
si no, sea μ el valor característico aproximado y
$\mathbf{w}' = (w'_1, \ldots, w'_{n-1})^t$ el vector característico aproximado.

Paso 7 Si $i \neq 1$, entonces para $k = 1, \ldots, i-1$ tome $w_k = w'_k$.

Paso 8 Tome $w_i = 0$.

Paso 9 Si $i \neq n$, entonces para $k = i+1, \ldots, n$ tome $w_k = w'_{k-1}$.

Paso 10 Para $k = 1, \ldots, n$

$$\text{tome } u_k = (\mu - \lambda)w_k + \left(\sum_{j=1}^{n} a_{ij}\, w_j\right) \frac{v_k}{v_i}.$$

(*Calcule el vector característico utilizando la ecuación (9.5).*)

Paso 11 SALIDA (μ, **u**); (*Procedimiento terminado exitosamente.*)
PARAR. ■

CONJUNTO DE EJERCICIOS 9.2

1. Encuentre las tres primeras iteraciones obtenidas con el método de la potencia aplicado a las siguientes matrices.

 a. $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (1, -1, 2)^t$.

 b. $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (-1, 0, 1)^t$.

 c. $\begin{bmatrix} 1 & -1 & 0 \\ -2 & 4 & -2 \\ 0 & -1 & 2 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (-1, 2, 1)^t$.

 d. $\begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (1, -2, 0, 3)^t$.

 e. $\begin{bmatrix} 5 & -2 & -\frac{1}{2} & \frac{3}{2} \\ -2 & 5 & \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} & 5 & -2 \\ \frac{3}{2} & -\frac{1}{2} & -2 & 5 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (1, 1, 0, -3)^t$.

 f. $\begin{bmatrix} -4 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -2 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 1 & 4 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (0, 0, 0, 1)^t$.

2. Repita el ejercicio 1 aplicando el método de la potencia inversa.

3. Encuentre las tres primeras iteraciones obtenidas con el método de la potencia simétrica aplicado a las siguientes matrices.

 a. $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (1, -1, 2)^t$.

 b. $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (-1, 0, 1)^t$.

 c. $\begin{bmatrix} 4.75 & 2.25 & -0.25 \\ 2.25 & 4.75 & 1.25 \\ -0.25 & 1.25 & 4.75 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (0, 1, 0)^t$.

 d. $\begin{bmatrix} 4 & 1 & -1 & 0 \\ 1 & 3 & -1 & 0 \\ -1 & -1 & 5 & 2 \\ 0 & 0 & 2 & 4 \end{bmatrix}$;
 Use $\mathbf{x}^{(0)} = (0, 1, 0, 0)^t$.

e. $\begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}$;

Use $\mathbf{x}^{(0)} = (1, 0, 0, 0)^t$.

f. $\begin{bmatrix} 5 & -2 & -\frac{1}{2} & \frac{3}{2} \\ -2 & 5 & \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} & 5 & -2 \\ \frac{3}{2} & -\frac{1}{2} & -2 & 5 \end{bmatrix}$;

Use $\mathbf{x}^{(0)} = (1, 1, 0, -3)^t$.

4. Desarrolle un algoritmo que incorpore el método de la potencia inversa en el de la potencia simétrica. Repita el ejercicio 3 usando el nuevo algoritmo.

5. Aplique el método de la potencia y la deflación de Wielandt para aproximar los dos valores característicos más dominantes de las matrices del ejercicio 1. Repita hasta obtener una tolerancia de 10^{-4} o hasta que el número de iteraciones sea mayor que 25.

6. Repita el ejercicio 5 usando el método Δ^2 de Aitken y el de las potencias, para el primer valor característico.

7. Aplique el método de la potencia simétrica para calcular el valor característico más grande (en valor absoluto) de las matrices del ejercicio 3. Repita hasta obtener una tolerancia de 10^{-4} o hasta que el número de iteraciones sea mayor que 25.

8. Repita el ejercicio 6 aplicando el método de la potencia inversa al primer valor característico.

9. Repita el ejercicio 7 aplicando el método de la potencia inversa.

10. **Técnica de anulación** Suponga que la matriz A de $n \times n$ tiene los valores característicos $\lambda_1, \ldots, \lambda_n$ ordenados por

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \cdots \geq |\lambda_n|,$$

con vectores característicos linealmente independientes $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}$.

a. Demuestre que, si el método de la potencia se aplica con un vector inicial $\mathbf{x}^{(0)}$ dado por

$$\mathbf{x}^{(0)} = \beta_2\mathbf{v}^{(2)} + \beta_3\mathbf{v}^{(3)} + \cdots + \beta_n\mathbf{v}^{(n)},$$

entonces la secuencia $\{\mu^{(m)}\}$ descrita en el algoritmo 9.1 convergirá en λ_2.

b. Demuestre que, para cualquier vector $\mathbf{x} = \sum_{i=1}^{n}\beta_i\mathbf{v}^{(i)}$, el vector $\mathbf{x}^{(0)} = (A - \lambda_1 I)\mathbf{x}$ satisface la propiedad dada en el inciso (a).

c. Obtenga una aproximación de λ_2 para las matrices del ejercicio 1.

d. Demuestre que este método puede continuarse para obtener λ_3 mediante $\mathbf{x}^{(0)} = (A - \lambda_2 I)(A - \lambda_1 I)\mathbf{x}$.

11. **Deflación de Hotelling** Suponga que el valor característico de mayor magnitud λ_1 y el vector característico asociado $\mathbf{v}^{(1)}$ se obtuvieron con la matriz simétrica A de $n \times n$. Demuestre que la matriz

$$B = A - \frac{\lambda_1}{(\mathbf{v}^{(1)})^t\mathbf{v}^{(1)}}\mathbf{v}^{(1)}(\mathbf{v}^{(1)})^t$$

tiene los mismos valores característicos $\lambda_2, \ldots, \lambda_n$ que A, excepto que B tiene el valor característico 0 con el vector característico $\mathbf{v}^{(1)}$, en vez del valor característico λ_1. Con este método de deflación calcule λ_2 para las matrices del ejercicio 3. En teoría, este método puede continuarse para obtener más valores característicos, pero pronto el error de redondeo hace inútil el esfuerzo.

12. Demuestre que el i-ésimo renglón de $B = A - \lambda_1\mathbf{v}^{(1)}\mathbf{x}^t$ es cero, donde λ_1 es el valor característico más grande de A en valor absoluto, $\mathbf{v}^{(1)}$ es el vector característico asociado de A para λ_1 y $\mathbf{x}$ es el vector definido en la ecuación (9.6).

13. De acuerdo con lo señalado en el ejercicio 11 de la sección 6.3 y en el ejercicio 11 de la sección 7.2, suponga que una especie de escarabajo tiene una longevidad de 4 años y que en el primer año una hembra tiene una supervivencia de 1/2, en el segundo año tiene una supervivencia

de 1/4 y en el tercer año una supervivencia de 1/8. Suponga, además, que una hembra pare, en promedio, dos hembras en el tercer año y cuatro hembras en el cuarto año. La matriz que describe la contribución de una hembra en un año a la población femenina en el año siguiente es

$$A=\begin{bmatrix} 0 & 0 & 2 & 4 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 \end{bmatrix},$$

donde una vez más el elemento del i-ésimo renglón y la j-ésima columna indica la contribución probabilística que una hembra de edad j hace a la población femenina de edad i en el siguiente año.

a. Aplique el teorema del círculo de Gerschgorin para determinar una región, dentro del plano complejo, que contenga todos los valores característicos de A.

b. Aplique el método de la potencia para determinar el valor característico dominante de la matriz y su vector característico asociado.

c. Use el algoritmo 9.4 para determinar cualesquiera valores característicos restantes y vectores característicos de A.

d. Calcule los valores característicos de A usando el polinomio característico de A y el método de Newton.

e. ¿Cuál es la predicción a largo plazo sobre la población de estos escarabajos?

14. Un sistema lineal dinámico puede representarse por medio de las ecuaciones

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x}(t) + B(t)\mathbf{u}(t), \qquad \mathbf{y}(t) = C(t)\mathbf{x}(t) + D(t)\mathbf{u}(t),$$

donde A es una matriz variable de $n \times n$, B es una matriz variable de $n \times r$, C es una matriz variable de $m \times n$, D es una matriz variable de $m \times r$, $\mathbf{x}$ es un vector variable n-dimensional, $\mathbf{y}$ es un vector variable m-dimensional y $\mathbf{u}$ es un vector variable r-dimensional. Para que el sistema sea estable, la matriz A debe tener todos sus valores característicos con parte real no positiva para toda t.

a. ¿Es el sistema estable si

$$A(t) = \begin{bmatrix} -1 & 2 & 0 \\ -2.5 & -7 & 4 \\ 0 & 0 & -5 \end{bmatrix}?$$

b. ¿Es el sistema estable si

$$A(t) = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -5 & 1 \\ -1 & -1 & -2 & -3 \end{bmatrix}?$$

15. La matriz tridiagonal $(m-1) \times (m-1)$

$$A = \begin{bmatrix} 1+2\alpha & -\alpha & 0 & \cdots & 0 \\ -\alpha & 1+2\alpha & -\alpha & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -\alpha \\ 0 & \cdots & 0 & -\alpha & 1+2\alpha \end{bmatrix}$$

interviene en el método de diferencias hacia atrás para resolver la ecuación del calor. (Véase la sección 12.2.) Para la estabilidad del método necesitamos que $\rho(A^{-1}) < 1$. Sea $m = 11$, aproxime $\rho(A^{-1})$ para los siguientes valores de α.

a. $\alpha = \frac{1}{4}$ b. $\alpha = \frac{1}{2}$ c. $\alpha = \frac{3}{4}$

¿Cuándo es el método estable?

16. Los valores característicos de la matriz A en el ejercicio 15 son

$$\lambda_i = 1 + 4\alpha\left(\text{sen}\,\frac{\pi i}{2m}\right)^2, \qquad \text{para } i = 1, \ldots, m-1.$$

Compare la aproximación del ejercicio 15 con el valor real de $\rho(A^{-1})$. Una vez más, ¿cuándo es el método estable?

17. Las matrices A y B de orden $(m-1) \times (m-1)$ dadas por

$$A = \begin{bmatrix} 1+\alpha & -\frac{\alpha}{2} & 0 & \cdots & 0 \\ -\frac{\alpha}{2} & 1+\alpha & -\frac{\alpha}{2} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -\frac{\alpha}{2} \\ 0 & \cdots & 0 & -\frac{\alpha}{2} & 1+\alpha \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 1-\alpha & \frac{\alpha}{2} & 0 & \cdots & 0 \\ \frac{\alpha}{2} & 1-\alpha & \frac{\alpha}{2} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{\alpha}{2} \\ 0 & \cdots & 0 & \frac{\alpha}{2} & 1-\alpha \end{bmatrix}$$

intervienen en el método de Crank-Nicolson para resolver la ecuación del calor (véase la sección 12.2). Con $m = 11$, aproxime $\rho(A^{-1}B)$ para los siguientes valores de α.

a. $\alpha = \frac{1}{4}$ b. $\alpha = \frac{1}{2}$ c. $\alpha = \frac{3}{4}$

9.3 Método de Householder

En la sección 9.4 usaremos el método QR para reducir una matriz tridiagonal simétrica a una matriz similar que sea casi diagonal. Las entradas en la diagonal de la matriz reducida son aproximaciones de los valores característicos de la matriz dada. En esta sección presentaremos un método diseñado por Alton Householder para reducir una matriz simétrica arbitraria a una matriz tridiagonal simétrica. Aunque existe una relación clara entre los problemas que estamos resolviendo en estas dos secciones, el método de Householder tiene amplia aplicación en otras áreas, además de la aproximación de valores característicos.

El método de Householder se usa para encontrar una matriz simétrica tridiagonal B que sea semejante a una matriz simétrica dada A. El teorema 9.10 implica que A es similar a una matriz diagonal D, ya que existe una matriz ortogonal Q con la propiedad de que $D = Q^{-1}AQ = Q^tAQ$. Como la matriz Q (y, en consecuencia, D) generalmente es difícil de calcular, el método de Householder ofrece una solución de compromiso. Una vez implantado este método podemos recurrir a métodos eficientes, como el algoritmo QR, para aproximar exactamente los valores característicos de la matriz tridiagonal simétrica resultante.

Definición 9.16 Sea $\mathbf{w} \in \mathbb{R}^n$ con $\mathbf{w}^t\mathbf{w} = 1$. La matriz de $n \times n$,

$$P = I - 2\mathbf{w}\mathbf{w}^t,$$

recibe el nombre de **transformación de Householder**. ■

Las transformaciones de Householder sirven para suprimir de manera selectiva bloques de elementos de vectores o columnas de matrices en una forma extremadamente estable respecto al error del redondeo. (Véase una explicación más amplia en [Wil2, pp. 152-162].) Las propiedades de estas transformaciones se dan en el siguiente teorema.

Teorema 9.17 Si $P = I - 2\mathbf{w}\mathbf{w}^t$ es una transformación de Householder, entonces P es simétrica y ortogonal; por tanto, $P^{-1} = P$. ■

Demostración Dado que

$$(\mathbf{w}\mathbf{w}^t)^t = (\mathbf{w}^t)^t\,\mathbf{w}^t = \mathbf{w}\mathbf{w}^t,$$

se deduce que

$$P^t = (I - 2\mathbf{w}\mathbf{w}^t)^t = I - 2\mathbf{w}\mathbf{w}^t = P.$$

Asimismo, como $\mathbf{w}^t\mathbf{w} = 1$,

$$\begin{aligned} PP^t &= (I - 2\mathbf{w}\mathbf{w}^t)(I - 2\mathbf{w}\mathbf{w}^t) = I - 2\mathbf{w}\mathbf{w}^t - 2\mathbf{w}\mathbf{w}^t + 4\mathbf{w}\mathbf{w}^t\mathbf{w}\mathbf{w}^t \\ &= I - 4\mathbf{w}\mathbf{w}^t + 4\mathbf{w}\mathbf{w}^t = I, \end{aligned}$$

por lo que

$$P^{-1} = P^t = P.$$

■ ■ ■

El método de Householder empieza con la determinación de una transformación $P^{(1)}$ tal que $A^{(2)} = P^{(1)}AP^{(1)}$ tiene

$$a_{j1}^{(2)} = 0, \quad \text{para cada } j = 3, 4, \ldots, n. \tag{9.7}$$

Por simetría, esto implica que $a_{1j}^{(2)} = 0$.

Se elige el vector $\mathbf{w} = (w_1, w_2, \ldots, w_n)^t$ de manera que $\mathbf{w}^t\mathbf{w} = 1$, lo que sostiene la ecuación (9.7), y en la matriz

$$A^{(2)} = P^{(1)}AP^{(1)} = (I - 2\mathbf{w}\mathbf{w}^t)\,A(I - 2\mathbf{w}\mathbf{w}^t),$$

tenemos $a_{11}^{(2)} = a_{11}$ y $a_{j1}^{(2)} = 0$, para cada $j = 3, 4, \ldots, n$. Esta elección impone n condiciones a las n incógnitas $w_1, \ldots, w_n$.

Al usar $w_1 = 0$ se garantiza que $a_{11}^{(2)} = a_{11}$. Queremos que

$$P^{(1)} = I - 2\mathbf{w}\mathbf{w}^t$$

satisfaga

$$P^{(1)}(a_{11}, a_{21}, a_{31}, \ldots, a_{n1})^t = (a_{11}, \alpha, 0, \ldots, 0)^t, \tag{9.8}$$

donde más tarde se seleccionará α. Para simplificar la notación, utilizamos

$$\hat{\mathbf{w}} = (w_2, w_3, \ldots, w_n)^t \in \mathbb{R}^{n-1}, \quad \hat{\mathbf{y}} = (a_{21}, a_{31}, \ldots, a_{n1})^t \in \mathbb{R}^{n-1},$$

y $\hat{P}$ como la $(n-1)\times(n-1)$ transformación de Householder

$$\hat{P} = I_{n-1} - 2\hat{\mathbf{w}}\hat{\mathbf{w}}^t.$$

La ecuación (9.8) se convierte entonces en

$$P^{(1)}\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \cdots\cdots 0 \\ 0 & \\ \vdots & \hat{P} \\ 0 & \end{bmatrix} \cdot \begin{bmatrix} a_{11} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} a_{11} \\ \hat{P}\hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} a_{11} \\ \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

con

$$\hat{P}\hat{\mathbf{y}} = (I_{n-1} - 2\hat{\mathbf{w}}\hat{\mathbf{w}}^t)\,\hat{\mathbf{y}} = \hat{\mathbf{y}} - 2(\hat{\mathbf{w}}^t\hat{\mathbf{y}})\hat{\mathbf{w}} = (\alpha, 0, \ldots, 0)^t. \tag{9.9}$$

Sea $r = \hat{\mathbf{w}}^t\hat{\mathbf{y}}$. Entonces

$$(\alpha, 0, \ldots, 0)^t = (a_{21} - 2rw_2, a_{31} - 2rw_3, \ldots, a_{n1} - 2rw_n)^t,$$

y podemos determinar todas las w_i si logramos determinar α y r. Al igualar los componentes, obtenemos

$$\alpha = a_{21} - 2rw_2$$

y

$$0 = a_{j1} - 2rw_j, \qquad \text{para cada } j = 3, \ldots, n.$$

Esto es,

$$2rw_2 = a_{21} - \alpha \tag{9.10}$$

y

$$2rw_j = a_{j1}, \qquad \text{para cada } j = 3, \ldots, n. \tag{9.11}$$

Al elevar al cuadrado ambos lados de las ecuaciones anteriores y al sumar, obtenemos

$$4r^2\sum_{j=2}^{n} w_j^2 = (a_{21} - \alpha)^2 + \sum_{j=3}^{n} a_{j1}^2.$$

Puesto que $\mathbf{w}^t\mathbf{w} = 1$ y $w_1 = 0$, tenemos $\sum_{j=2}^{n} w_j^2 = 1$ y

$$4r^2 = \sum_{j=2}^{n} a_{j1}^2 - 2\alpha a_{21} + \alpha^2. \tag{9.12}$$

Con base en la ecuación (9.9) y en el hecho de que P es ortogonal, tenemos

$$\alpha^2 = (\alpha, 0, \ldots, 0)(\alpha, 0, \ldots, 0)^t = (\hat{P}\hat{\mathbf{y}})^t\hat{P}\hat{\mathbf{y}} = \hat{\mathbf{y}}^t\hat{P}^t\hat{P}\hat{\mathbf{y}} = \hat{\mathbf{y}}^t\hat{\mathbf{y}}.$$

Por tanto,

$$\alpha^2 = \sum_{j=2}^{n} a_{j1}^2,$$

que, al ser sustituida en (9.12), nos da:

$$2r^2 = \sum_{j=2}^{n} a_{j1}^2 - \alpha a_{21}.$$

Para asegurarnos de que $2r^2 = 0$ sólo si $a_{21} = a_{31} = \cdots = a_{n1} = 0$, elegimos el signo, de modo que

$$\alpha = -\operatorname{sgn}(a_{21}) \left(\sum_{j=2}^{n} a_{j1}^2 \right)^{1/2},$$

lo que implica que

$$2r^2 = \sum_{j=2}^{n} a_{j1}^2 + |a_{21}| \left(\sum_{j=2}^{n} a_{j1}^2 \right)^{1/2}.$$

Con estas opciones de α y $2r^2$ resolvemos las ecuaciones (9.10) y (9.11) para obtener

$$w_2 = \frac{a_{21} - \alpha}{2r} \quad \text{y} \quad w_j = \frac{a_{j1}}{2r}, \quad \text{para cada } j = 3, \ldots, n.$$

Para resumir la elección de $P^{(1)}$, tenemos

$$\alpha = -\operatorname{sgn}(a_{21}) \left(\sum_{j=2}^{n} a_{j1}^2 \right)^{1/2},$$

$$r = \left(\frac{1}{2}\alpha^2 - \frac{1}{2}a_{21}\alpha \right)^{1/2},$$

$$w_1 = 0,$$

$$w_2 = \frac{a_{21} - \alpha}{2r}$$

y

$$w_j = \frac{a_{j1}}{2r}, \quad \text{para cada } j = 3, \ldots, n.$$

Con esta opción elegida,

$$A^{(2)} = P^{(1)}AP^{(1)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & 0 & \cdots & 0 \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}.$$

Una vez que hemos encontrado $P^{(1)}$ y $A^{(2)}$, el proceso se repite para $k = 2, 3, \ldots, n-2$ como sigue:

$$\alpha = -\operatorname{sgn}(a_{k+1,k}^{(k)}) \left(\sum_{j=k+1}^{n} (a_{jk}^{(k)})^2 \right)^{1/2},$$

$$r = \left(\frac{1}{2}\alpha^2 - \frac{1}{2}\alpha a_{k+1,k}^{(k)} \right)^{1/2},$$

$$w_1^{(k)} = w_2^{(k)} = \ldots = w_k^{(k)} = 0,$$

$$w_{k+1}^{(k)} = \frac{a_{k+1,k}^{(k)} - \alpha}{2r},$$

$$w_j^{(k)} = \frac{a_{jk}^{(k)}}{2r}, \quad \text{para cada } j = k+2, k+3, \ldots, n,$$

$$P^{(k)} = I - 2\mathbf{w}^{(k)} \cdot (\mathbf{w}^{(k)})^t$$

y

$$A^{(k+1)} = P^{(k)} A^{(k)} P^{(k)},$$

donde

$$A^{(k+1)} = \begin{bmatrix} a_{11}^{(k+1)} & a_{12}^{(k+1)} & 0 & \cdots & \cdots & \cdots & 0 \\ a_{21}^{(k+1)} & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 & \cdots & 0 \\ \vdots & & \ddots & a_{k+1,k}^{(k+1)} & a_{k+1,k+1}^{(k+1)} & a_{k+1,k+2}^{(k+1)} \cdots & a_{k+1,n}^{(k+1)} \\ \vdots & & & 0 & \vdots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{bmatrix}.$$

Continuando de este modo se forma la matriz tridiagonal y simétrica $A^{(n-1)}$, donde

$$A^{(n-1)} = P^{(n-2)} P^{(n-3)} \ldots P^{(1)} A P^{(1)} \ldots P^{(n-3)} P^{(n-2)},$$

EJEMPLO 1 La matriz 4×4

$$A = \begin{bmatrix} 4 & 1 & -2 & 2 \\ 1 & 2 & 0 & 1 \\ -2 & 0 & 3 & -2 \\ 2 & 1 & -2 & -1 \end{bmatrix}$$

es simétrica. Para la primera aplicación de una transformación de Householder,

$$\alpha = -(1)\left(\sum_{j=2}^{4} a_{j1}^2\right)^{1/2} = -3, \quad r = \left(\frac{1}{2}(-3)^2 - \frac{1}{2}(1)(-3)\right)^{1/2} = \sqrt{6},$$

$$\mathbf{w} = \left(0, \frac{\sqrt{6}}{3}, -\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{6}\right),$$

$$P^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - 2\left(\frac{\sqrt{6}}{6}\right)^2 \begin{bmatrix} 0 \\ 2 \\ -1 \\ 1 \end{bmatrix} \cdot (0, 2, -1, 1)$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

y

$$A^{(2)} = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & 1 & \frac{4}{3} \\ 0 & 1 & \frac{5}{3} & -\frac{4}{3} \\ 0 & \frac{4}{3} & -\frac{4}{3} & -1 \end{bmatrix}.$$

Continuamos con la segunda iteración para obtener

$$\alpha = -\frac{5}{3}, \qquad r = \frac{2\sqrt{5}}{3}, \qquad \mathbf{w} = \left(0, 0, 2\sqrt{5}, \frac{\sqrt{5}}{5}\right)^t,$$

$$P^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & 0 & -\frac{4}{5} & \frac{3}{5} \end{bmatrix}$$

y la matriz tridiagonal simétrica es

$$A^{(3)} = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & \frac{10}{3} & -\frac{5}{3} & 0 \\ 0 & -\frac{5}{3} & -\frac{33}{25} & \frac{68}{75} \\ 0 & 0 & \frac{68}{75} & \frac{149}{75} \end{bmatrix}. \qquad \blacksquare$$

En el algoritmo 9.5 se ejecuta el método de Householder que acabamos de describir, aunque se evitan las multiplicaciones matriciales reales.

ALGORITMO 9.5

Método de Householder

Para obtener una matriz tridiagonal simétrica $A^{(n-1)}$ similar a la matriz simétrica $A = A^{(1)}$, se construyen las siguientes matrices $A^{(2)}, A^{(3)}, \ldots, A^{(n-1)}$, donde $A^{(k)} = (a_{ij}^{(k)})$ para cualquier $k = 1, 2, \ldots, n-1$:

ENTRADA dimensión n; matriz A.

SALIDA $A^{(n-1)}$. (*En cada paso podemos sobreescribir A.*)

Paso 1 Para $k = 1, 2, \ldots, n-2$, haga los pasos 2-14.

Paso 2 Tome

$$q = \sum_{j=k+1}^{n} (a_{jk}^{(k)})^2.$$

Paso 3 Si $a_{k+1,k}^{(k)} = 0$, entonces tome $\alpha = -q^{1/2}$

$$\text{si no, tome } \alpha = -\frac{q^{1/2} a_{k+1,k}^{(k)}}{|a_{k+1,k}^{(k)}|}.$$

Paso 4 Tome $RSQ = \alpha^2 - \alpha a_{k+1,k}^{(k)}$. (*Nota*: $RSQ = 2r^2$)

Paso 5 Tome $v_k = 0$; (*Nota*: $v_1 = \cdots = v_{k-1} = 0$, *pero no se necesitan.*)
$v_{k+1} = a_{k+1,k}^{(k)} - \alpha$;
Para $j = k+2, \ldots, n$ tome $v_j = a_{jk}^{(k)}$.

$$\left(\textit{Nota}: \mathbf{w} = \left(\frac{1}{\sqrt{2 RSQ}}\right)\mathbf{v} = \frac{1}{2r}\mathbf{v}.\right)$$

Paso 6 Para $j = k, k+1, \ldots, n$ tome $u_j = \left(\frac{1}{RSQ}\right)\sum_{i=k+1}^{n} a_{ji}^{(k)} v_i$.

$$\left(\textit{Nota}: \mathbf{u} = \left(\frac{1}{RSQ}\right)A^{(k)}\mathbf{v} = \frac{1}{2r^2}A^{(k)}\mathbf{v} = \frac{1}{r}A^{(k)}\mathbf{w}.\right)$$

Paso 7 Tome $PROD = \sum_{i=k+1}^{n} v_i u_i$.

$$\left(\textit{Nota}: PROD = \mathbf{v}^t\mathbf{u} = \frac{1}{2r^2}\mathbf{v}^t A^{(k)}\mathbf{v}.\right)$$

Paso 8 Para $j = k, k+1, \ldots, n$ tome $z_j = u_j - \left(\frac{PROD}{2RSQ}\right)v_j$.

$$\left(\textit{Nota}: \mathbf{z} = \mathbf{u} - \frac{1}{2RSQ}\mathbf{v}^t\mathbf{u}\mathbf{v} = \mathbf{u} - \frac{1}{4r^2}\mathbf{v}^t\mathbf{u}\mathbf{v}\right.$$

$$\left. = \mathbf{u} - \mathbf{w}\mathbf{w}^t\mathbf{u} = \frac{1}{r}A^{(k)}\mathbf{w} - \mathbf{w}\mathbf{w}^t\frac{1}{r}A^{(k)}\mathbf{w}.\right)$$

Paso 9 Para $l = k+1, k+2, \ldots, n-1$, haga los pasos 10 y 11.
(*Nota: calcule* $A^{(k+1)} = A^{(k)} - \mathbf{v}\mathbf{z}^t - \mathbf{z}\mathbf{v}^t = (I - 2\mathbf{w}\mathbf{w}^t)A^{(k)}(I - 2\mathbf{w}\mathbf{w}^t)$.)

Paso 10 Para $j = l + 1, \ldots, n$ tome
$a_{jl}^{(k+1)} = a_{jl}^{(k)} - v_l z_j - v_j z_l$;
$a_{lj}^{(k+1)} = a_{jl}^{(k+1)}$.

Paso 11 Tome $a_{ll}^{(k+1)} = a_{ll}^{(k)} - 2v_l z_l$.

Paso 12 Tome $a_{nn}^{(k+1)} = a_{nn}^{(k)} - 2v_n z_n$.

Paso 13 Para $j = k + 2, \ldots, n$ tome $a_{kj}^{(k+1)} = a_{jk}^{(k+1)} = 0$.

Paso 14 Tome $a_{k+1,k}^{(k+1)} = a_{k+1,k}^{(k)} - v_{k+1} z_k$;
$a_{k,k+1}^{(k+1)} = a_{k+1,k}^{(k)}$.
(*Nota: los otros elementos de $A^{(k+1)}$ son los mismos de $A^{(k)}$.*)

Paso 15 SALIDA ($A^{(n-1)}$);
(*El proceso se terminó. $A^{(n-1)}$ es simétrico, tridiagonal y similar a A.*)
PARAR. ■

Si queremos aplicar el algoritmo de Householder a una matriz arbitraria de $n \times n$, debemos hacer la siguiente modificación para compensar una posible falta de simetría.

Paso 6 Para $j = 1, 2, \ldots, n$ tome $u_j = \frac{1}{RSQ} \sum_{i=k+1}^{n} a_{ji}^{(k)} v_i$;

$$y_j = \frac{1}{RSQ} \sum_{i=k+1}^{n} a_{ij}^{(k)} v_i.$$

Paso 8 Para $j = 1, 2, \ldots, n$ tome $z_j = u_j - \frac{PROD}{RSQ} v_j$.

Paso 9 Para $l = k + 1, k + 2, \ldots, n$, haga los pasos 10 y 11.

Paso 10 Para $j = 1, 2, \ldots, k$ tome $a_{jl}^{(k+1)} = a_{jl}^{(k)} - z_j v_l$;
$a_{lj}^{(k+1)} = a_{lj}^{(k)} - y_j v_l$.

Paso 11 Para $j = k + 1, \ldots, n$ tome $a_{jl}^{(k+1)} = a_{jl}^{(k)} - z_j v_l - y_l v_j$.

Una vez modificados los pasos anteriores, se suprimen los pasos 12 a 14 y la salida $A^{(n-1)}$.

La matriz resultante $A^{(n-1)}$ no será tridiagonal, a menos que la matriz original A sea simétrica, pero sí tendrá sólo elementos cero debajo de la subdiagonal inferior. Una matriz de este tipo recibe el nombre de *Hessenberg superior*. Es decir, $H = (h_{ij})$ es **Hessenberg superior** si $h_{ij} = 0$ para cualquier $i \geq j + 2$.

En la siguiente sección veremos cómo podemos aplicar el algoritmo QR a $A^{(n-1)}$ para determinar sus valores característicos, que son iguales a los de la matriz original A.

CONJUNTO DE EJERCICIOS 9.3

1. Aplique el método de Householder para poner las siguientes matrices en forma tridiagonal.

a. $\begin{bmatrix} 12 & 10 & 4 \\ 10 & 8 & -5 \\ 4 & -5 & 3 \end{bmatrix}$ b. $\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$

c. $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ d. $\begin{bmatrix} 4.75 & 2.25 & -0.25 \\ 2.25 & 4.75 & 1.25 \\ -0.25 & 1.25 & 4.75 \end{bmatrix}$

2. Aplique el método de Householder para poner las siguientes matrices en forma tridiagonal.

a. $\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}$ b. $\begin{bmatrix} 5 & -2 & -0.5 & 1.5 \\ -2 & 5 & 1.5 & -0.5 \\ -0.5 & 1.5 & 5 & -2 \\ 1.5 & -0.5 & -2 & 5 \end{bmatrix}$

c. $\begin{bmatrix} 8 & 0.25 & 0.5 & 2 & -1 \\ 0.25 & -4 & 0 & 1 & 2 \\ 0.5 & 0 & 5 & 0.75 & -1 \\ 2 & 1 & 0.75 & 5 & -0.5 \\ -1 & 2 & -1 & -0.5 & 6 \end{bmatrix}$

d. $\begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 3 & 0 & -2 & 0 \\ -1 & 0 & 4 & 2 & 1 \\ 0 & -2 & 2 & 8 & 3 \\ -0 & 0 & 1 & 3 & 9 \end{bmatrix}$

3. Modifique el algoritmo 9.5 de Householder para calcular matrices superiores de Hessenberg similares con las siguientes matrices no simétricas.

a. $\begin{bmatrix} 2 & -1 & 3 \\ 2 & 0 & 1 \\ -2 & 1 & 4 \end{bmatrix}$ b. $\begin{bmatrix} -1 & 2 & 3 \\ 2 & 3 & -2 \\ 3 & 1 & -1 \end{bmatrix}$

c. $\begin{bmatrix} 5 & -2 & -3 & 4 \\ 0 & 4 & 2 & -1 \\ 1 & 3 & -5 & 2 \\ -1 & 4 & 0 & 3 \end{bmatrix}$ d. $\begin{bmatrix} 4 & -1 & -1 & -1 \\ -1 & 4 & 0 & -1 \\ -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & 4 \end{bmatrix}$

9.4 Algoritmo QR

Los métodos de deflación expuestos en la sección 9.2 generalmente no son adecuados para calcular todos los valores característicos de una matriz, debido al crecimiento del error de redondeo. En esta sección estudiaremos el algoritmo QR, que es una técnica de reducción matricial que permite determinar simultáneamente todos los valores característicos de una matriz simétrica.

Para aplicar el método QR, partimos de una matriz simétrica en forma tridiagonal; es decir, las únicas entradas no nulas de la matriz están en la diagonal o en las subdiagonales directamente arriba o debajo de la diagonal. Si ésta no es la forma de la matriz simétrica, el primer paso consiste en aplicar el método de Householder para calcular una matriz simétrica tridiagonal similar a la matriz dada.

En esta sección supondremos que la matriz simétrica cuyos valores característicos vamos a calcular es tridiagonal. Si con A denotamos una matriz de este tipo, podremos simplificar un poco la notación marcando los elementos de A así:

$$A = \begin{bmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & a_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{bmatrix}. \tag{9.13}$$

Si $b_2 = 0$ o $b_n = 0$, entonces la matriz 1×1 $[a_1]$ o bien $[a_n]$ produce inmediatamente un valor característico a_1 o a_n de A.

Cuando $b_j = 0$ para alguna j, donde $2 < j < n$, el problema se puede minimizar considerando, en vez de A, las matrices más pequeñas

$$\begin{bmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & a_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{j-1} \\ 0 & \cdots & 0 & b_{j-1} & a_{j-1} \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} a_j & b_{j+1} & 0 & \cdots & 0 \\ b_{j+1} & a_{j+1} & b_{j+2} & \ddots & \vdots \\ 0 & b_{j+2} & a_{j+2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{bmatrix} \tag{9.14}$$

Si ninguna b_j es cero, el método QR forma una sucesión de las matrices $A = A^{(1)}, A^{(2)}, A^{(3)}, \ldots,$ así:

1. $A^{(1)} = A$ se factoriza como un producto $A^{(1)} = Q^{(1)}R^{(1)}$, donde $Q^{(1)}$ es ortogonal y $R^{(1)}$ es triangular superior.
2. $A^{(2)}$ se define como $A^{(2)} = R^{(1)}Q^{(1)}$.

En general, $A^{(i)}$ se factoriza como un producto $A^{(i)} = Q^{(i)}R^{(i)}$ de una matriz ortogonal $Q^{(i)}$ y de una matriz triangular superior $R^{(i)}$. Después, definimos $A^{(i+1)}$ como el producto de $R^{(i)}$ y $Q^{(i)}$ en la dirección inversa $A^{(i+1)} = R^{(i)}Q^{(i)}$. Dado que $Q^{(i)}$ es ortogonal, $R^{(i)} = Q^{(i)t} A^{(i)}$ y

$$A^{(i+1)} = R^{(i)} Q^{(i)} = (Q^{(i)t} A^{(i)}) Q^{(i)} = Q^{(i)t} A^{(i)} Q^{(i)}. \tag{9.15}$$

Y $A^{(i+1)}$ es simétrica con los mismos valores característicos que $A^{(i)}$. Por la forma en que definimos $R^{(i)}$ y $Q^{(i)}$, también podemos garantizar que $A^{(i+1)}$ es tridiagonal.

Continuando por inducción, $A^{(i+1)}$ tiene los mismos valores característicos que la matriz original A. El éxito del procedimiento se debe al hecho de que $A^{(i+1)}$ tiende a una matriz diagonal con los valores característicos de A a lo largo de la diagonal.

Si queremos describir la construcción de las matrices de factor $Q^{(i)}$ y $R^{(i)}$, necesitamos manejar el concepto de *matriz de rotación*.

Definición 9.18 Una **matriz de rotación** P difiere de la matriz identidad en cuatro elementos como máximo. Estos tienen la forma

$$p_{ii} = p_{jj} = \cos\theta \quad \text{y} \quad p_{ij} = -p_{ji} = \operatorname{sen}\theta,$$

para alguna θ e $i \neq j$. ■

Es fácil demostrar (véase el ejercicio 6) que, con cualquier matriz de rotación P, la matriz AP difiere de A sólo en la i-ésima y j-ésima columnas y la matriz PA difiere de A sólo en el i-ésimo y j-ésimo renglones. Para cualquier $i \neq j$, podemos elegir el ángulo θ de modo que el producto PA tenga un elemento cero para $(PA)_{ij}$. Además, toda matriz de rotación P es ortogonal, porque la definición implica que $PP^t = I$.

La factorización de $A^{(1)}$ en $A^{(1)} = Q^{(1)}R^{(1)}$ utiliza un producto de $n - 1$ matrices de rotación de este tipo para construir

$$R^{(1)} = P_n P_{n-1} \cdots P_2 A^{(1)}.$$

Primero establecemos que la matriz de rotación P_2 tenga

$$p_{11} = p_{22} = \cos\theta_2 \quad \text{y} \quad p_{12} = -p_{21} = \operatorname{sen}\theta_2,$$

donde

$$\operatorname{sen}\theta_2 = \frac{b_2}{\sqrt{b_2^2 + a_1^2}} \quad \text{y} \quad \cos\theta_2 = \frac{a_1}{\sqrt{b_2^2 + a_1^2}};$$

entonces, la matriz

$$A_2^{(1)} = P_2 A^{(1)}$$

tiene un cero en la posición (2, 1), esto es, en el segundo renglón y en la primera columna, ya que el elemento (2, 1) de $A_2^{(1)}$ es

$$(-\operatorname{sen}\theta_2)a_1 + (\cos\theta_2)b_2 = \frac{-b_2 a_1}{\sqrt{b_2^2 + a_1^2}} + \frac{a_1 b_2}{\sqrt{b_2^2 + a_1^2}} = 0.$$

Como la multiplicación $P_2 A^{(1)}$ afecta a los renglones 1 y 2 de $A^{(1)}$, la nueva matriz no necesariamente conserva los elementos cero en las posiciones (1, 3), (1, 4), ..., y (1, n). Sin embargo, $A^{(1)}$ es tridiagonal y, por tanto, los elementos (1, 4), ..., (1, n) de $A_2^{(1)}$ son cero. Sólo el elemento (1, 3), el del primer renglón y tercera columna, puede hacerse distinto de cero.

En términos generales, seleccionamos la matriz P_k de modo que el (k, $k - 1$)-ésimo elemento de $A_k^{(1)} = P_k A_{k-1}^{(1)}$ sea cero, lo cual hace que el ($k - 1$, $k + 1$)-elemento se convierta en uno distinto de cero. La matriz $A_k^{(1)}$ tiene la forma

$$A_k^{(1)} = \begin{bmatrix} z_1 & q_1 & r_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 0 & z_{k-1} & q_{k-1} & r_{k-1} & \ddots & \vdots \\ \vdots & & \ddots & 0 & x_k & y_k & 0 & \cdots & 0 \\ \vdots & & & \ddots & b_{k+1} & a_{k+1} & b_{k+2} & \ddots & 0 \\ \vdots & & & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & & & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & b_n & a_n \end{bmatrix}$$

y P_{k+1} tiene la forma

$$P_{k+1} = \left[\begin{array}{c|cc|c} I_{k-1} & O & & O \\ \hline & c_{k+1} & s_{k+1} & \\ O & & & O \\ & -s_{k+1} & c_{k+1} & \\ \hline O & O & & I_{n-k-1} \end{array}\right] \begin{array}{l} \\ \leftarrow \text{renglón } k \\ \\ \\ \\ \end{array} \qquad (9.16)$$

$\uparrow$ columna k

donde O denota la matriz nula con la dimensión adecuada.

Elegimos las constantes $c_{k+1} = \cos\theta_{k+1}$ y $s_{k+1} = \operatorname{sen}\theta_{k+1}$ en P_{k+1}, de modo que el $(k + 1, k)$ elemento de $A_{k+1}^{(1)}$ sea cero; es decir, $s_{k+1}x_k - c_{k+1}b_{k+1} = 0$. Puesto que $c_{k+1}^2 + s_{k+1}^2 = 1$, la solución a esta ecuación es

$$s_{k+1} = \frac{b_{k+1}}{\sqrt{b_{k+1}^2 + x_k^2}} \quad \text{y} \quad c_{k+1} = \frac{x_k}{\sqrt{b_{k+1}^2 + x_k^2}}$$

y $A_{k+1}^{(2)}$ tiene la forma

$$A_k^{(1)} = \begin{bmatrix} z_1 & q_1 & r_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 0 & z_k & q_k & r_k & \ddots & \vdots \\ \vdots & & \ddots & 0 & x_{k+1} & y_{k+1} & 0 & \cdots & 0 \\ \vdots & & & \ddots & b_{k+2} & a_{k+2} & b_{k+3} & \ddots & 0 \\ \vdots & & & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & & & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & b_n & a_n \end{bmatrix}$$

Al proseguir con esta construcción en la sucesión $P_2, \ldots, P_n$ obtenemos la matriz triangular superior

$$R^{(1)} = A_n^{(1)} = \begin{bmatrix} z_1 & q_1 & r_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & r_{n-2} \\ \vdots & & & \ddots & z_{n-1} & q_{n-1} \\ 0 & \cdots & \cdots & \cdots & 0 & x_n \end{bmatrix}.$$

La otra mitad de la factorización QR es la matriz

$$Q^{(1)} = P_2^t P_3^t \cdots P_n^t,$$

porque la ortogonalidad de las matrices de rotación implica que

$$Q^{(1)}R^{(1)} = (P_2^t P_3^t \cdots P_n^t) \cdot (P_n \cdots P_3 P_2)A^{(1)} = A^{(1)}.$$

La matriz $Q^{(1)}$ es ortogonal porque

$$(Q^{(1)})^t Q^{(1)} = (P_2^t P_3^t \cdots P_n^t)^t (P_2^t P_3^t \cdots P_n^t) = (P_n \cdots P_3 P_2) \cdot (P_2^t P_3^t \cdots P_n^t) = I.$$

Además, $Q^{(1)}$ es una matriz Hessenberg superior. Para comprobar por qué lo es, el lector puede seguir los pasos de los ejercicios 7 y 8.

En consecuencia, $A^{(2)} = R^{(1)}Q^{(1)}$ es también una matriz Hessenberg superior, ya que la multiplicación de $Q^{(1)}$ de la izquierda por la matriz triangular superior $R^{(1)}$ no influye en los elementos del triángulo inferior. Lo anterior implica que $A^{(2)}$ efectivamente es tridiagonal, pues ya sabemos que es simétrica.

Los elementos situados fuera de la diagonal de $A^{(2)}$ generalmente serán más pequeños que los correspondientes de $A^{(1)}$, así que $A^{(2)}$ se acerca más a ser una matriz diagonal que $A^{(1)}$. El proceso se repite para construir $A^{(3)}, A^{(4)}, \ldots$.

Si los valores característicos de A tienen módulos distintos, con $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, entonces la rapidez de convergencia del elemento $b_{j+1}^{(i+1)}$ a 0 en la matriz $A^{(i+1)}$ depende del cociente $|\lambda_{j+1}/\lambda_j|$ (véase [Fr]). La rapidez de convergencia de $b_{j+1}^{(i+1)}$ a 0 determina la razón con la que el elemento $a_j^{(i+1)}$ converge al j-ésimo valor característico λ_j. Así, la rapidez de convergencia puede ser lenta si $|\lambda_{j+1}/\lambda_j|$ está cerca de la unidad.

Si queremos acelerar esta convergencia, se utiliza un método de desplazamiento semejante al empleado con el método de las potencias inversas de la sección 9.2. Se selecciona una constante s cercana a un valor característico de A, con lo cual se modifica la factorización de la ecuación (9.15) para escoger $Q^{(i)}$ y $R^{(i)}$, de modo que

$$A^{(i)} - sI = Q^{(i)} R^{(i)} \tag{9.17}$$

y, equivalentemente, podemos definir que la matriz $A^{(i+1)}$ sea

$$A^{(i+1)} = R^{(i)} Q^{(i)} + sI. \tag{9.18}$$

Con esta modificación, la rapidez de convergencia de $b_{j+1}^{(i+1)}$ a cero depende de la razón $|(\lambda_{j+1} - s)/(\lambda_j - s)|$, la cual puede ocasionar una mejora significativa en la rapidez original de convergencia de $a_j^{(i+1)}$ en λ_j si s está cerca de λ_{j+1}, pero no de λ_j.

En el listado del algoritmo QR, cambiamos s en cada paso para que, cuando A tenga los valores característicos del módulo definido, $b_n^{(i+1)}$ converja a cero más rápido que $b_j^{(i+1)}$ para cualquier entero j menor que n. Cuando $b_n^{(i+1)}$ es suficientemente pequeño, suponemos que $\lambda_n \approx a_n^{(i+1)}$, eliminamos los n-ésimos renglón y columna de la matriz y procedemos en la misma forma para obtener una aproximación de λ_{n-1}. El proceso continúa hasta determinar una aproximación para cada valor característico.

El algoritmo incorpora el método de desplazamiento al seleccionar en el i-ésimo paso la constante de deplazamiento s_i, donde s_i es el valor característico más cercano a $a_n^{(i)}$ de la matriz

$$E^{(i)} = \begin{bmatrix} a_{n-1}^{(i)} & b_n^{(i)} \\ b_n^{(i)} & a_n^{(i)} \end{bmatrix}.$$

Este desplazamiento traduce los valores característicos de A por un factor s_i. Con esta técnica de desplazamiento, la convergencia suele ser cúbica. (Véase [WR, p. 270].) El algoritmo acumula estos desplazamientos hasta que $b_n^{(i+1)} \approx 0$ y luego agrega los desplazamientos $a_n^{(i+1)}$ para aproximar el valor característico λ_n.

Si A tiene los valores característicos del mismo módulo, $b_j^{(i+1)}$ puede tender a cero para alguna $j \neq n$ con mayor rapidez que $b_n^{(i+1)}$. En este caso, el método de separación de matrices descrito en (9.14) puede servir para convertir el problema en uno que contenga un par de matrices de orden reducido.

EJEMPLO 1 Sea

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} a_1^{(1)} & b_2^{(1)} & 0 \\ b_2^{(1)} & a_2^{(1)} & b_3^{(1)} \\ 0 & b_3^{(1)} & a_3^{(1)} \end{bmatrix}.$$

Para obtener el parámetro de aceleración del desplazamiento necesitamos los valores característicos de

$$\begin{bmatrix} a_2^{(1)} & b_3^{(1)} \\ b_3^{(1)} & a_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix},$$

que son $\mu_1 = 4$ y $\mu_2 = 2$. La elección del valor característico más cercano a $a_3^{(1)} = 3$ es arbitraria y escogemos $\mu_2 = 2$ y lo cambiamos por esta cantidad. Así, $s_1 = 2$ y

$$\begin{bmatrix} d_1 & b_2^{(1)} & 0 \\ b_2^{(1)} & d_2 & b_3^{(1)} \\ 0 & b_3^{(1)} & d_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Al continuar los cálculos, obtenemos

$$x_1 = 1, \quad y_1 = 1, \quad z_1 = \sqrt{2}, \quad c_2 = \frac{\sqrt{2}}{2}, \quad s_2 = \frac{\sqrt{2}}{2},$$

$$q_1 = \sqrt{2}, \quad x_2 = 0, \quad r_1 = \frac{\sqrt{2}}{2} \quad \text{y} \quad y_2 = \frac{\sqrt{2}}{2}.$$

Por tanto,

$$A_2^{(1)} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \sqrt{2} \\ 0 & 1 & 1 \end{bmatrix}.$$

Asimismo,

$$z_2 = 1, \quad c_3 = 0, \quad s_3 = 1, \quad q_2 = 1 \quad \text{y} \quad x_3 = -\frac{\sqrt{2}}{2}.$$

En consecuencia,

$$R^{(1)} = A_3^{(1)} = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 1 & 1 \\ 0 & 0 & -\frac{\sqrt{2}}{2} \end{bmatrix}.$$

Para calcular $A^{(2)}$, tenemos

$$z_3 = -\frac{\sqrt{2}}{2}, \quad a_1^{(2)} = 2, \quad b_2^{(2)} = \frac{\sqrt{2}}{2}, \quad a_2^{(2)} = 1, \quad b_3^{(2)} = -\frac{\sqrt{2}}{2} \quad \text{y} \quad a_3^{(2)} = 0,$$

por lo que

$$A^{(2)} = R^{(1)} Q^{(1)} = \begin{bmatrix} 2 & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & 1 & -\frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & 0 \end{bmatrix}.$$

Se termina una iteración del método QR. Puesto que ni $b_2^{(2)} = \sqrt{2}/2$ ni $b_3^{(2)} = -\sqrt{2}/2$ son pequeños, se realizará otra iteración del algoritmo QR. Para esta iteración, calculamos los valores característicos $\frac{1}{2} \pm \frac{1}{2}\sqrt{3}$ de la matriz

$$\begin{bmatrix} a_2^{(2)} & b_3^{(2)} \\ b_3^{(2)} & a_3^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

y elegimos $s_2 = \frac{1}{2} - \frac{1}{2}\sqrt{3}$, el valor característico más cercano a $a_3^{(2)} = 0$. Completamos los cálculos para obtener

$$A^{(3)} = \begin{bmatrix} 2.6720277 & 0.37597448 & 0 \\ 0.37597448 & 1.4736080 & 0.030396964 \\ 0 & 0.030396964 & -0.047559530 \end{bmatrix}.$$

Si $b_3^{(3)} = 0.030396964$ es suficientemente pequeño, entonces la aproximación al valor característico λ_3 es 1.5864151, la suma de $a_3^{(3)}$ y $s_1 + s_2 = 2 + (1 - \sqrt{3})/2$. Eliminando el tercer renglón y la tercera columna obtenemos

$$A^{(3)} = \begin{bmatrix} 2.6720277 & 0.37597448 \\ 0.37597448 & 1.4736080 \end{bmatrix},$$

que tiene los valores característicos $\mu_1 = 2.7802140$ y $\mu_2 = 1.3654218$. Al agregar el factor de desplazamiento obtenemos las aproximaciones

$$\lambda_1 \approx 4.4141886 \quad \text{y} \quad \lambda_2 \approx 2.9993964.$$

Puesto que los valores característicos de la matriz A son 4.41420, 3.00000 y 1.58579, el método QR dio cuatro cifras significativas de exactitud en sólo dos iteraciones. ∎

El algoritmo 9.6 ejecuta el método QR de este modo.

ALGORITMO 9.6

Algoritmo QR

Para obtener los valores característicos de la matriz tridiagonal simétrica de $n \times n$

$$A \equiv A_1 = \begin{bmatrix} a_1^{(1)} & b_2^{(1)} & 0 & \cdots & 0 \\ b_2^{(1)} & a_2^{(1)} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n^{(1)} \\ 0 & \cdots & 0 & b_n^{(1)} & a_n^{(1)} \end{bmatrix}$$

ENTRADA n; $a_1^{(1)}, \ldots, a_n^{(1)}$, $b_2^{(1)}, \ldots, b_n^{(1)}$; tolerancia TOL; número máximo de iteraciones M.

SALIDA valores característicos de A, una división recomendada de A o un mensaje de que excedió el número máximo de iteraciones.

Paso 1 Tome $k = 1$;
$SHIFT = 0$. (*Shift acumulado.*)

Paso 2 Mientras $k \leq M$, haga los pasos 3-19.
(Los pasos 3-7 prueban el éxito.)

Paso 3 Si $|b_n^{(k)}| \leq TOL$, entonces tome $\lambda = a_n^{(k)} + SHIFT$;
SALIDA (λ);
tome $n = n - 1$.

Paso 4 Si $|b_2^{(k)}| \leq TOL$, entonces tome $\lambda = a_1^{(k)} + SHIFT$;
SALIDA (λ);
tome $n = n - 1$;
$a_1^{(k)} = a_2^{(k)}$;
para $j = 2, \ldots, n$
tome $a_j^{(k)} = a_{j+1}^{(k)}$;
$b_j^{(k)} = b_{j+1}^{(k)}$.

Paso 5 Si $n = 0$, entonces
PARAR.

Paso 6 Si $n = 1$, entonces
tome $\lambda = a_1^{(k)} + SHIFT$;
SALIDA (λ);
PARAR.

Paso 7 Para $j = 3, \ldots, n-1$
si $|b_j^{(k)}| \leq TOL$, entonces
SALIDA('dividido en' $a_1^{(k)}, \ldots, a_{j-1}^{(k)}, b_2^{(k)}, \ldots, b_{j-1}^{(k)}$,
'y',
$a_j^{(k)}, \ldots, a_n^{(k)}, b_{j+1}^{(k)}, \ldots, b_n^{(k)}, SHIFT$);
PARAR.

Paso 8 (*Calcule shift.*)
Tome $b = -(a_{n-1}^{(k)} + a_n^{(k)})$;
$c = a_n^{(k)} a_{n-1}^{(k)} - [b_n^{(k)}]^2$;
$d = (b^2 - 4c)^{1/2}$.

Paso 9 Si $b > 0$, entonces tome $\mu_1 = -2c/(b + d)$;
$\mu_2 = -(b + d)/2$;
si no, tome $\mu_1 = (d - b)/2$;
$\mu_2 = 2c/(d - b)$.

Paso 10 Si $n = 2$, entonces tome $\lambda_1 = \mu_1 + SHIFT$;
$\lambda_2 = \mu_2 + SHIFT$;
SALIDA (λ_1, λ_2);
PARAR.

Paso 11 Elija s tal que $|s - a_n^{(k)}| = \min\{|\mu_1 - a_n^{(k)}|, |\mu_2 - a_n^{(k)}|\}$.

Paso 12 (*Shift acumulado.*)
Tome $SHIFT = SHIFT + s$.

Paso 13 (*Shift realizado.*)
Para $j = 1, \ldots, n$, tome $d_j = a_j^{(k)} - s$.

Paso 14 (*Pasos 14 y 15 calculan $R^{(k)}$.*)
Tome $x_1 = d_1$;
$y_1 = b_2$.

Paso 15 Para $j = 2, \ldots, n$

tome $z_{j-1} = \left\{x_{j-1}^2 + \left[b_j^{(k)}\right]^2\right\}^{1/2}$;

$c_j = \dfrac{x_{j-1}}{z_{j-1}}$;

$s_j = \dfrac{b_j^{(k)}}{z_{j-1}}$;

$q_{j-1} = c_j y_{j-1} + s_j d_j$;
$x_j = -s_j y_{j-1} + c_j d_j$;

Si $j \neq n$, entonces tome $r_{j-1} = s_j b_{j+1}^{(k)}$;

$y_j = c_j b_{j+1}^{(k)}$.

$\left(A_j^{(k)} = P_j A_{j-1}^{(k)}\right.$ *apenas se ha calculado y* $\left.R^{(k)} = A_n^{(k)}.\right)$

Paso 16 (*Pasos* 16-18 *calculan* $A^{(k+1)}$).
Tome $z_n = x_n$;
$a_1^{(k+1)} = s_2 q_1 + c_2 z_1$;
$b_2^{(k+1)} = s_2 z_2$.

Paso 17 Para $j = 2, 3, \ldots, n-1$
tome $a_j^{(k+1)} = s_{j+1} q_j + c_j c_{j+1} z_j$;
$b_{j+1}^{(k+1)} = s_{j+1} z_{j+1}$.

Paso 18 Tome $a_n^{(k+1)} = c_n z_n$.

Paso 19 Tome $k = k + 1$.

Paso 20 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Podemos aplicar un procedimiento parecido para obtener las aproximaciones a los valores característicos de una matriz no simétrica de $n \times n$. Primero, reducimos la matriz a una matriz Hessenberg superior similar H, usando el algoritmo de Householder para matrices no simétricas.

El proceso de factorización QR adopta la siguiente forma. Primero

$$H \equiv H^{(1)} = Q^{(1)} R^{(1)}. \tag{9.19}$$

Después, definimos $H^{(2)}$ mediante

$$H^{(2)} = R^{(1)} Q^{(1)} \tag{9.20}$$

y la factorizamos en

$$H^{(2)} = Q^{(2)} R^{(2)}. \tag{9.21}$$

El método de factorización prosigue con el mismo objetivo que el algoritmo QR. Seleccionamos las matrices para introducir ceros en elementos apropiados de la matriz y aplicamos un procedimiento de desplazamiento semejante al del método QR. Sin embargo, en el caso de las matrices no simétricas, el desplazamiento es un poco más complicado, porque pueden presentarse valores característicos complejos con el mismo módulo. El proceso de desplazamiento modifica los cálculos de las ecuaciones (9.19), (9.20) y (9.21) para obtener el método doble QR, $H^{(1)} - s_1 I = Q^{(1)} R^{(1)}$, $H^{(2)} = R^{(1)} Q^{(1)} + s_1 I$, $H^{(2)} - s_2 I = Q^{(2)} R^{(2)}$ y $H^{(3)} = R^{(2)} Q^{(2)} + s_2 I$, donde s_1 y s_2 son conjugadas complejas y $H^{(1)}, H^{(2)}, \ldots$, son matrices Hessenberg superiores reales.

Una descripción completa del método QR se encuentra en [Wil2]. En [WR] encontrará algoritmos y programas detallados de este método y de la mayor parte de otros de uso común. Recomendamos al lector consultar esas obras si el método que hemos expuesto no le da resultados satisfactorios.

El método QR puede aplicarse en forma tal que produzca los vectores característicos de una matriz y también sus valores característicos; en cambio, el algoritmo 9.6 no fue diseñado para hacer eso. Si además de los valores característicos se necesitan los vectores característicos de una matriz simétrica, sugerimos utilizar el método de las potencias inversas después de usar los algoritmos 9.5 y 9.6, o bien, aplicar una técnica más potente, como las que se describen en [WR], que son métodos ideados expresamente para este fin.

CONJUNTO DE EJERCICIOS 9.4

1. Aplique dos iteraciones del algoritmo QR a las siguientes matrices.

a. $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ **b.** $\begin{bmatrix} 3 & 1 & 0 \\ 1 & 4 & 2 \\ 0 & 2 & 1 \end{bmatrix}$

c. $\begin{bmatrix} 4 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ **d.** $\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & -1 & 0 \\ 0 & -1 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}$

e. $\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -3 & -1 & 0 \\ 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix}$ **f.** $\begin{bmatrix} 0.5 & 0.25 & 0 & 0 \\ 0.25 & 0.8 & 0.4 & 0 \\ 0 & 0.4 & 0.6 & 0.1 \\ 0 & 0 & 0.1 & 1 \end{bmatrix}$

2. Use el algoritmo QR para determinar, con una exactitud de 10^{-5}, los valores característicos de las siguientes matrices.

a. $\begin{bmatrix} 2 & -1 & 0 \\ -1 & -1 & -2 \\ 0 & -2 & 3 \end{bmatrix}$ **b.** $\begin{bmatrix} 3 & 1 & 0 \\ 1 & 4 & 2 \\ 0 & 2 & 3 \end{bmatrix}$

c. $\begin{bmatrix} 4 & 2 & 0 & 0 & 0 \\ 2 & 4 & 2 & 0 & 0 \\ 0 & 2 & 4 & 2 & 0 \\ 0 & 0 & 2 & 4 & 2 \\ 0 & 0 & 0 & 2 & 4 \end{bmatrix}$ **d.** $\begin{bmatrix} 5 & -1 & 0 & 0 & 0 \\ -1 & 4.5 & 0.2 & 0 & 0 \\ 0 & 0.2 & 1 & -0.4 & 0 \\ 0 & 0 & -0.4 & 3 & 1 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix}$

3. Use el algoritmo QR para determinar, con una exactitud de 10^{-5}, todos los valores característicos de las matrices del ejercicio 1.

4. Aplique el método de las potencias inversas para determinar, con una exactitud de 10^{-5}, los vectores característicos de las matrices del ejercicio 1.

5. **a.** Demuestre que la matriz de rotación $\begin{bmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{bmatrix}$ aplicada al vector $\mathbf{x} = (x_1, x_2)^t$ tiene el efecto geométrico de girar $\mathbf{x}$ a través del ángulo θ sin que cambie su magnitud respecto a $\|\cdot\|_2$.

b. Demuestre que la magnitud de $\mathbf{x}$ respecto a $\|\cdot\|_\infty$ puede cambiarse mediante una matriz de rotación.

6. Sea P la matriz de rotación con $p_{ii} = p_{jj} = \cos\theta$ y $p_{ij} = -p_{ji} = \operatorname{sen}\theta$, para $j < i$. Demuestre que para toda matriz A de $n \times n$:

$$(AP)_{pq} = \begin{cases} a_{pq}, & \text{si } q \neq i, j, \\ (\cos\theta)a_{pj} + (\text{sen}\,\theta)a_{pi}, & \text{si } q = j, \\ (\cos\theta)a_{pi} - (\text{sen}\,\theta)a_{pj}, & \text{si } q = i. \end{cases}$$

$$(PA)_{pq} = \begin{cases} a_{pq}, & \text{si } p \neq i, j, \\ (\cos\theta)a_{jq} - (\text{sen}\,\theta)a_{iq}, & \text{si } p = j, \\ (\text{sen}\,\theta)a_{jq} + (\cos\theta)a_{iq}, & \text{si } p = i. \end{cases}$$

7. Demuestre que el producto de una matriz triangular superior (a la izquierda) y una matriz Hessenberg superior produce una matriz Hessenberg superior.

8. Denote con P_k a una matriz de rotación de la forma dada en (9.16).

   a. Demuestre que $P_2^t P_3^t$ difiere de una matriz triangular superior solamente en las posiciones (2, 1) y (3, 2) como máximo.

   b. Suponga que $P_2^t P_3^t \cdots P_k^t$ difiere de una matriz triangular superior sólo en las posiciones (2, 1), (3, 2), …, $(k, k-1)$ como máximo. Demuestre que $P_2^t P_3^t \cdots P_k^t P_{k+1}^t$ difiere de una matriz triangular superior sólo en las posiciones (2, 1), (3, 2), …, $(k, k-1)$, $(k+1, k)$ como máximo.

   c. Demuestre que la matriz $P_2^t P_3^t \cdots P_n^t$ es Hessenberg superior.

9. El **método de Jacobi** para una matriz simétrica A está descrito por

$$A_1 = A,$$
$$A_2 = P_1 A_1 P_1^t$$

y, en general, por

$$A_{i+1} = P_i A_i P_i^t.$$

La matriz A_{i+1} tiende a ser diagonal, donde P_i es una matriz de rotación escogida para eliminar un elemento grande fuera de la diagonal de A_i. Si queremos hacer cero a $a_{j,k}$ y $a_{k,j}$, donde $j \neq k$, entonces $a_{jj} \neq a_{kk}$ si

$$(P_i)_{jj} = (P_i)_{kk} = \sqrt{\frac{1}{2}\left(1 + \frac{b}{\sqrt{c^2 + b^2}}\right)},$$

$$(P_i)_{kj} = \frac{c}{2(P_i)_{jj}\sqrt{c^2 + b^2}} = -(P_i)_{jk},$$

donde

$$c = 2a_{jk}\,\text{sgn}(a_{jj} - a_{kk}) \quad \text{y} \quad b = |a_{jj} - a_{kk}|,$$

o bien, $a_{jj} = a_{kk}$,

$$(P_i)_{jj} = (P_i)_{kk} = \frac{\sqrt{2}}{2}$$

y

$$(P_i)_{kj} = -(P_i)_{jk} = \frac{\sqrt{2}}{2}.$$

Desarrolle un algoritmo para implantar o ejecutar el método de Jacobi usando $a_{21} = 0$. Después, tome como cero $a_{31}, a_{32}, a_{41}, a_{42}, a_{43}, \ldots, a_{n,1}, \ldots a_{n,n-1}$. Repita esto hasta calcular una matriz A_k con

$$\sum_{i=k}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}^{(k)}|$$

suficientemente pequeña. Entonces, aproxime los valores característicos de A mediante los elementos diagonales de A_k.

10. Repita el ejercicio 3 aplicando el método de Jacobi.

11. En el primer ejemplo de este capítulo, hay que resolver el sistema lineal $A\mathbf{w} = -0.04(\rho/p)\lambda\mathbf{w}$ para $\mathbf{w}$ y λ a fin de aproximar los valores característicos λ_k del sistema de Strum-Liouville.

a. Calcule los cuatro valores característicos $\mu_1, \ldots, \mu_4$ de la matriz

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

con una exactitud de 10^{-5}.

b. Aproxime los valores característicos $\lambda_1, \ldots, \lambda_4$ del sistema en función de ρ y p.

12. La matriz tridiagonal $(m-1) \times (m-1)$

$$A = \begin{bmatrix} 1-2\alpha & \alpha & 0 & \cdots & 0 \\ \alpha & 1-2\alpha & \alpha & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \alpha \\ 0 & \cdots & 0 & \alpha & 1-2\alpha \end{bmatrix}$$

interviene en el método de diferencias hacia adelante con que se resuelve la ecuación del calor (véase la sección 12.2). Para lograr la estabilidad del método necesitamos $\rho(A) < 1$. Sea $m = 11$, aproxime los valores característicos de A en cada uno de los siguientes incisos.

a. $\alpha = \frac{1}{4}$ **b.** $\alpha = \frac{1}{2}$ **c.** $\alpha = \frac{3}{4}$

¿Cuándo es estable el método?

13. Los valores característicos de la matriz A del ejercicio 12 son

$$\lambda_i = 1 - 4\alpha\left(\operatorname{sen} \frac{\pi i}{2m}\right)^2, \quad \text{para } i = 1, \ldots, m-1.$$

Compare la aproximación del ejercicio 12 con el valor característico real. Una vez más, ¿cuándo el método es estable?

9.5 Reseña de métodos y de software

En este capítulo estudiamos la aproximación de los valores y vectores característicos. Los círculos de Gerschgorin ofrecen una aproximación rudimentaria a la ubicación que tienen los valores característicos de una matriz. Podemos utilizar el método de la potencia para

calcular el valor dominante y un vector característico asociado de una matriz arbitraria A. Si A es simétrica, el método de la potencia simétrica da una convergencia más rápida al valor característico dominante y el vector característico asociado. El método de la potencia inversa permite obtener el valor característico más cercano a determinado valor y un vector característico asociado. Con este método, a menudo se refina un valor característico aproximado y se calcula un vector característico, una vez obtenido el valor característico mediante otra técnica.

Los métodos de deflación, como la deflación de Wielandt, producen otros valores característicos, una vez conocido el valor característico dominante. Estos métodos, por ser vulnerables al error del redondeo, se aplican sólo si se requieren unos cuantos valores característicos. El método de la potencia inversa debe emplearse para mejorar la exactitud de los valores característicos aproximados obtenidos con un método de deflación.

Los métodos basados en las transformaciones de similitud, como el de Householder, sirven para convertir una matriz simétrica en una matriz similar tridiagonal (o Hessenberg superior, en caso de que la matriz no sea simétrica). Técnicas como el método QR pueden aplicarse después a la matriz tridiagonal (de Hessenberg superior) para obtener aproximaciones a los valores característicos. Los vectores característicos asociados pueden calcularse mediante un método iterativo (por ejemplo, el de la potencia inversa) aplicado a los valores característicos obtenidos con el método QR. Nuestro estudio se limitó a las matrices simétricas, y explicamos este método para calcular los valores característicos sólo en caso de simetría.

Las subrutinas de las bibliotecas IMSL y NAG se basan en las contenidas en EISPACK y LAPACK, que ya explicamos en la sección 1.4. En términos generales, las subrutinas transforman una matriz en la forma apropiada para aplicar el método QR o una de sus modificaciones, como el método QL. Las subrutinas aproximan todos los valores característicos, pudiendo aproximar también un vector asociado con cada valor característico. Hay rutinas especiales que calculan todos los valores característicos de un intervalo o región, o bien, sólo el valor característico más grande o más pequeño. También existen subrutinas para aproximar la exactitud del valor característico y la sensibilidad del proceso ante el error de redondeo.

La rutina SGEBAL de LAPACK prepara una matriz no simétrica real A para procesamiento posterior. Trata de usar las matrices de permutación para convertir A en una forma triangular superior de bloque. Las transformaciones de similitud sirven para equilibrar los renglones y columnas de la norma. La rutina SGEHRD puede utilizarse después para convertir A en una matriz Hessenberg superior similar, H. Esta matriz puede, entonces, ser reducida mediante SHSEQR en la forma STS^t, de Schur, donde S es ortogonal y la diagonal de T contiene los valores característicos de A. STREVC puede servir, entonces, para obtener los vectores característicos correspondientes.

La rutina SSYTRD de LAPACK sirve para reducir una matriz simétrica real A en una matriz tridiagonal semejante mediante el método de Householder. La rutina SSTEQR utiliza un algoritmo QR implícitamente desplazado para obtener los valores y vectores característicos de A.

La subrutina EVLRG de IMSL produce todos los valores de A por orden creciente de magnitud. Primero, equilibra la matriz A por medio de una versión de la rutina BALANC de EISPACK, de modo que las sumas de las magnitudes de los elementos de los renglones y de las columnas sean aproximadamente iguales. Ello origina mayor estabilidad en los cálculos subsecuentes. EVLRG realiza, entonces, transformaciones ortogonales de similitud, como en el método de Householder, para reducir A a una matriz Hessenberg superior semejante. Esta parte se parece a la subrutina ORTHES de EISPACK. Finalmente, para ob-

tener todos los valores característicos, se realiza el algoritmo QR de desplazamiento. Esta parte se parece a la subrutina HQR de EISPACK. La subrutina EVCRG de IMSL es igual que EVRLG, salvo que se calculan los vectores característicos correspondientes. La subrutina EVLSF calcula los valores característicos de la matriz simétrica real A. Primero, mediante una modificación de la rutina TRED2 de EISPACK, se reduce la matriz A a la forma tridiagonal. Después, se calculan los valores característicos aplicando una modificación de la rutina IMTQL2 de EISPACK, la cual es una variante del método QR denominada método implícito QL. La subrutina EVCSF es igual que EVLSF, excepto que se calculan además los vectores característicos. Finalmente, EVLRH y EVCRH calculan todos los valores característicos de la matriz Hessenberg superior A y, además, EVCRH calcula los vectores característicos. Estas subrutinas se basan en las subrutinas HQR y HQR2, respectivamente, de EISPACK.

La biblioteca NAG cuenta con subrutinas que se basan en las rutinas de EISPACK. La subrutina F02EBF calcula los valores característicos de una matriz real y, opcionalmente, los vectores característicos. La subrutina F02AGF es igual que F02AFF, excepto que también se calculan los vectores característicos. Primero se equilibra la matriz y luego se reduce a la forma Hessenberg superior para aplicar el método QR. Si sólo se requieren los valores característicos, el algoritmo emplea un método QR Hessenberg para calcularlos; si también se requieren los vectores característicos, se emplea una factorización Schur. La subrutina F02FAF se emplea con una matriz simétrica real. Los valores característicos, que son reales, se calculan por orden creciente de magnitud. Si además se necesitan los vectores característicos, puede aplicarse la subrutina F02ABF. En uno y otro caso, la matriz se reduce a la forma tridiagonal aplicando el método de Householder y después se calculan los valores característicos mediante el algoritmo QR. La subrutina F08FEF ejecuta directamente el algoritmo de Householder con matrices simétricas para producir una matriz simétrica tridiagonal similar. La biblioteca NAG cuenta también con rutinas para equilibrar directamente las matrices reales, recuperando los vectores característicos, si primero se equilibra la matriz, y efectuando otras operaciones con tipos especiales de matrices.

El procedimiento de Maple `Eigenvals(A)` calcula los valores característicos de A equilibrando primero A y luego transformándola a la forma tridiagonal o a la forma Hessenberg superior. Después, se aplica el método QR para obtener todos los valores y vectores característicos. Al igual que en el algoritmo 9.6, la forma tridiagonal se emplea para una matriz simétrica.

El procedimiento de MATLAB `eig` calcula los valores característicos y, opcionalmente, los vectores característicos de A usando las rutinas EISPACK. Usa BALANC para equilibrar la matriz, ORTHES para transformar la matriz a una forma superior de Hessenberg y por último, una rutina HQR2 modificada para calcular los valores característicos y, opcionalmente, los vectores característicos de una matriz real superior de Hessenberg mediante el método QR. MATLAB también tiene un procedimiento `eigs` que calcula un número elegido de valores y vectores característicos. El procedimiento `eigs` se basa en el método con reinicio implícito de Arnoldi creado por Sorensen [Sor]. El paquete de software ARPAC [ARP] de Netlib para resolver problemas de valores característicos grandes y esparcidos se basa también en este último método, el cual usa subespacios de Krylov y determina una serie de tales subespacios, que converge a un subespacio que contiene los valores característicos.

Los libros de Wilkinson [Wil2] y de Wilkinson y Reinsch [WR] son clásicos en el estudio de los problemas referentes a los valores característicos. Stewart [St] es otra buena fuente de información acerca del problema general y Parlett [Par] explica el problema simétrico. Un estudio del problema no simétrico se encuentra en Saad [Sa1].

CAPÍTULO

10

Soluciones numéricas de sistemas de ecuaciones no lineales

• • •

La presión que se requiere para sumir un objeto grande y pesado en un suelo blando y homogéneo situado arriba de un terreno de base dura, puede predecirse mediante la presión que se requiere para introducir objetos más pequeños en el mismo terreno. Específicamente, la presión p para sumir una placa circular de radio r a una distancia d en un terreno blando, donde el terreno de base sólida se halla a una distancia $D > d$ debajo de la superficie, puede aproximarse con una ecuación de la forma

$$p = k_1 e^{k_2 r} + k_3 r,$$

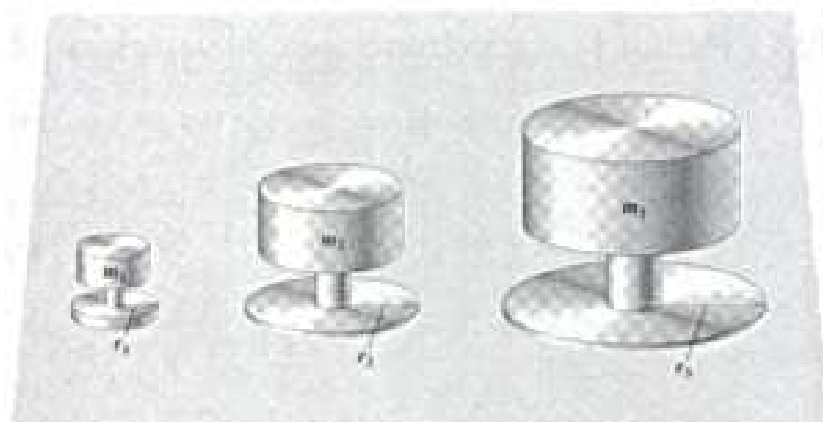
donde k_1, k_2 y k_3 son constantes que dependen de d y de la consistencia del suelo, pero no del radio de la placa.

Si queremos determinar el tamaño mínimo de la placa necesario para sostener una gran carga, metemos tres placas pequeñas con radios distintos a la misma distancia; en la figura siguiente se muestran las cargas requeridas para esta maniobra.

Esto genera las tres ecuaciones no lineales

$$m_1 = k_1 e^{k_2 r_1} + k_3 r_1,$$
$$m_2 = k_1 e^{k_2 r_2} + k_3 r_2,$$
$$m_3 = k_1 e^{k_2 r_3} + k_3 r_3,$$

en las tres incógnitas k_1, k_2 y k_3. Para resolver los sistemas de ecuaciones cuando éstas son no lineales, por lo general se necitan los



métodos de aproximación numérica. El ejercicio 10 de la sección 10.2 se refiere a una aplicación del tipo que acabamos de describir ahora.

El problema de resolver un sistema de ecuaciones no lineales se evita en la medida de lo posible; por lo general, se aproxima el sistema no lineal mediante un sistema de ecuaciones lineales. Cuando esto no es satisfactorio, el problema debe enfrentarse en forma directa. El punto de vista más directo consiste en adaptar los métodos del capítulo 2, que aproximan las soluciones de una sola ecuación no lineal en una variable, la cual se reemplaza por un problema vectorial que incorpora todas las variables.

La principal herramienta en el capítulo 2 fue el método de Newton, una técnica que por lo general converge en forma cuadrática. Ésta es la primera técnica que modificaremos para resolver sistemas de ecuaciones no lineales. El método de Newton, modificado para sistemas de ecuaciones, es muy costoso, de modo que en la sección 10.3 describiremos la forma de usar un método modificado de la secante para obtener aproximaciones más fácilmente, aunque perdiendo la convergencia extremadamente rápida proporcionada por el método de Newton.

La sección 10.4 describe el método del descenso más rápido. Sólo es linealmente convergente, pero no necesita las aproximaciones iniciales tan precisas como las técnicas de convergencia más rápida. Con frecuencia se usa para determinar una buena aproximación inicial para el método de Newton o una de sus modificaciones.

En la sección 10.5 daremos una introducción a los métodos de continuación, que usan un parámetro para pasar de un problema con una solución fácilmente determinada a la solución del problema no lineal original.

La mayoría de las demostraciones de los resultados teóricos de este capítulo se omite, por requerir métodos que normalmente se estudian en cálculo avanzado. Una buena referencia general de consulta para este tema es el libro de Ortega titulado *Numerical Analysis-A Second Course* [Or2]. Una obra más completa es [OR].

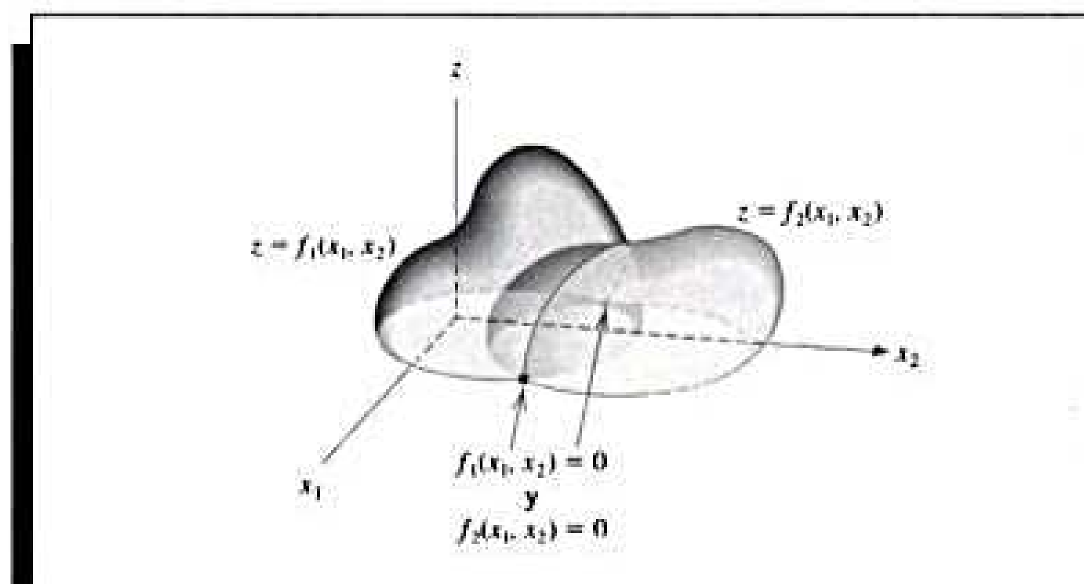
10.1 Puntos fijos para funciones de varias variables

Un sistema de ecuaciones no lineales tiene la forma

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0,\\ f_2(x_1, x_2, \ldots, x_n) &= 0,\\ \vdots \qquad & \quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned} \tag{10.1}$$

donde podemos considerar a toda función f_i como un mapeo de un vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ del espacio n-dimensional $\mathbb{R}^n$ en la recta real $\mathbb{R}$. En la figura 10.1 se muestra una representación geométrica de un sistema no lineal cuando $n = 2$.

Figura 10.1



Este sistema de n ecuaciones no lineales con n incógnitas puede representarse también mediante la definición de una función $\mathbf{F}$, mapeando $\mathbb{R}^n$ en $\mathbb{R}^n$ por medio de

$$\mathbf{F}(x_1, x_2, \ldots, x_n) = (f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots, f_n(x_1, x_2, \ldots, x_n))^t.$$

Si se emplea una notación vectorial para representar las variables $x_1, x_2, \ldots, x_n$, el sistema (10.1) adopta la forma

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}. \tag{10.2}$$

Las funciones $(f_1, f_2, \ldots, f_n)$ son, entonces, las **funciones coordenadas de F**.

EJEMPLO 1 El sistema no lineal de 3×3

$$3x_1 - \cos(x_2 x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 81(x_2 + 0.1)^2 + \text{sen}\, x_3 + 1.06 = 0,$$

$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

puede ponerse en la forma (10.2) definiendo las tres funciones f_1, f_2 y f_3 de $\mathbb{R}^3$ en $\mathbb{R}$ como

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - \frac{1}{2},$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \text{sen}\, x_3 + 1.06,$$

$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3},$$

y $\mathbf{F}$ de $\mathbb{R}^3 \to \mathbb{R}^3$ por

$$\begin{aligned}\mathbf{F}(\mathbf{x}) &= \mathbf{F}(x_1, x_2, x_3)\\ &= (f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3), f_3(x_1, x_2, x_3))^t\\ &= \left(3x_1 - \cos(x_2x_3) - \frac{1}{2}, x_1^2 - 81(x_2 + 0.1)^2 + \text{sen}\, x_3 + 1.06,\right.\\ &\qquad \left. e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3}\right)^t.\end{aligned}$$

■

Antes de explicar la manera de resolver un sistema de la forma (10.1) o (10.2), necesitamos algunos resultados referentes a la continuidad y diferenciabilidad de las funciones de $\mathbb{R}^n$ a $\mathbb{R}^n$. Aunque podríamos presentar este estudio directamente (véase el ejercicio 10), recurrimos a un método alterno que nos permite explicar los conceptos de límites y continuidad, más difíciles desde el punto de vista teórico, a partir de las funciones de $\mathbb{R}^n$ en $\mathbb{R}$.

Definición 10.1 Sea f una función definida en el conjunto $D \subset \mathbb{R}^n$ en $\mathbb{R}$. Se dice que la función f tiene el límite L en $\mathbf{x}_0$, y se escribe

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = L,$$

si, dado un número cualquiera $\varepsilon > 0$, existe un número $\delta > 0$ con la propiedad de que

$$|f(\mathbf{x}) - L| < \varepsilon,$$

siempre que $\mathbf{x} \in D$ y

$$0 < \|\mathbf{x} - \mathbf{x}_0\| < \delta.$$

■

La existencia de un límite es independiente de la norma vectorial que se utilice (véase la sección 7.1). Se puede usar cualquier norma conveniente para satisfacer la condición de esta definición. El valor específico de δ es independiente de la norma.

Definición 10.2 Sea f una función del conjunto $D \subset \mathbb{R}^n$ en $\mathbb{R}$. La función f es **continua** en $\mathbf{x}_0 \in D$ siempre que exista $\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x})$ y que

$$\lim_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0).$$

Más aún, f es **continua en un conjunto** D si f es continua en todos los puntos de D. Este concepto se expresa escribiendo $f \in C(D)$. ■

Ahora ya podemos definir los conceptos de límite y continuidad para las funciones de $\mathbb{R}^n$ en $\mathbb{R}^n$ considerando las funciones coordenadas de $\mathbb{R}^n$ en $\mathbb{R}$.

Definición 10.3 Sea $\mathbf{F}$ una función de $D \subset \mathbb{R}^n$ en $\mathbb{R}^n$ de la forma

$$\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))^t,$$

donde f_i se mapea de $\mathbb{R}^n$ en $\mathbb{R}$ para toda i. Definimos

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{L} = (L_1, L_2, \ldots, L_n)^t$$

si y sólo si $\lim_{\mathbf{x}\to\mathbf{x}_0} f_i(\mathbf{x}) = L_i$ para toda $i = 1, 2, \ldots, n$. ■

La función $\mathbf{F}$ es *continua* en $\mathbf{x}_0 \in D$ siempre y cuando $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x})$ exista y $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_0)$. Además, $\mathbf{F}$ es continua en el conjunto D si lo es en cada $\mathbf{x}$ de D. Este concepto lo expresamos escribiendo $F \in C(D)$.

En el caso de funciones de $\mathbb{R}$ en $\mathbb{R}$, a menudo podemos probar la continuidad demostrando que la función es diferenciable (véase el teorema 1.6). Aunque este teorema se generaliza a funciones de varias variables, la derivada (o la derivada total) de una función de varias variables resulta muy compleja y no la explicaremos aquí. En lugar de ello enunciaremos el siguiente teorema que relaciona la continuidad de una función de n variables en un punto con las derivadas parciales de la función en ese punto.

Teorema 10.4 Sea f una función de $D \subset \mathbb{R}^n$ en $\mathbb{R}$ y $\mathbf{x}_0 \in D$. Si existen las constantes $\delta > 0$ y $K > 0$ con

$$\left| \frac{\partial f(\mathbf{x})}{\partial x_j} \right| \leq K, \text{ para cada } j = 1, 2, \ldots, n,$$

siempre que $\|\mathbf{x} - \mathbf{x}_0\| < \delta$ y $\mathbf{x} \in D$, entonces f es continua en $\mathbf{x}_0$. ■

En el capítulo 2 desarrollamos un proceso iterativo para resolver una ecuación $f(x) = 0$, transformándola primero en la forma $x = g(x)$. Por definición, la función g tiene puntos fijos precisamente en las soluciones de la ecuación original. Para el caso de las funciones de $\mathbb{R}^n$ en $\mathbb{R}^n$, se investigará un procedimiento semejante.

Definición 10.5 Una función $\mathbf{G}$ de $D \subset \mathbb{R}^n$ en $\mathbb{R}^n$ tiene un **punto fijo** en $\mathbf{p} \in D$ si $\mathbf{G}(\mathbf{p}) = \mathbf{p}$. ■

El siguiente teorema generaliza al teorema del punto fijo 2.3 para el caso n-dimensional. Este teorema es un caso especial del teorema de mapeo de contracción, y su demostración viene en [Or2, p. 153].

Teorema 10.6 Sea $D = \{(x_1, x_2, \ldots, x_n)^t | a_i \leq x_i \leq b_i$ para toda $i = 1, 2, \ldots, n\}$ para algún conjunto de constantes $a_1, a_2, \ldots, a_n$ y $b_1, b_2, \ldots, b_n$. Supongamos que **G** es una función continua de $D \subset \mathbb{R}^n$ en $\mathbb{R}^n$ con la propiedad de que $\mathbf{G}(\mathbf{x}) \in D$ siempre que $\mathbf{x} \in D$. Entonces **G** tiene un punto fijo en D.

Supongamos, además, que G tiene derivadas parciales continuas y que existe una constante $K < 1$ con

$$\left| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \right| \leq \frac{K}{n}, \quad \text{siempre que } \mathbf{x} \in D,$$

para toda $j = 1, 2, \ldots, n$ y toda función componente g_i. Entonces la sucesión $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ definida por una $\mathbf{x}^{(0)}$ seleccionada en D arbitrariamente y generada por

$$\mathbf{x}^{(k)} = G(\mathbf{x}^{(k-1)}), \quad \text{para cada } k \geq 1,$$

converge en el único punto fijo $\mathbf{p} \in D$ y

$$\|\mathbf{x}^{(k)} - \mathbf{p}\|_\infty \leq \frac{K^k}{1-k} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty. \tag{10.3}$$

■

EJEMPLO 2 Considere el sistema no lineal del ejemplo 1 dado por

$$3x_1 - \cos(x_2 x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

Si resolvemos la i-ésima ecuación para x_i, el sistema se transforma en un problema de punto fijo

$$x_1 = \frac{1}{3}\cos(x_2 x_3) + \frac{1}{6},$$

$$x_2 = \frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1, \tag{10.4}$$

$$x_3 = -\frac{1}{20} e^{-x_1 x_2} - \frac{10\pi - 3}{60}.$$

Supongamos que $\mathbf{G}: \mathbb{R}^3 \to \mathbb{R}^3$ está definida por $\mathbf{G}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), g_3(\mathbf{x}))^t$, donde

$$g_1(x_1, x_2, x_3) = \frac{1}{3}\cos(x_2 x_3) + \frac{1}{6},$$

$$g_2(x_1, x_2, x_3) = \frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1,$$

$$g_3(x_1, x_2, x_3) = -\frac{1}{20} e^{-x_1 x_2} - \frac{10\pi - 3}{60}.$$

Utilizaremos los teoremas 10.4 y 10.6 para demostrar que **G** tiene un punto fijo único en

$$D = \{(x_1, x_2, x_3)^t \mid -1 \leq x_i \leq 1, \quad \text{para cada } i = 1, 2, 3\}.$$

Para $\mathbf{x} = (x_1, x_2, x_3)^t$ en D,

$$|g_1(x_1, x_2, x_3)| \leq \frac{1}{3}|\cos(x_2x_3)| + \frac{1}{6} \leq 0.50,$$

$$|g_2(x_1, x_2, x_3)| = \left|\frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1\right| \leq \frac{1}{9}\sqrt{1 + \operatorname{sen} 1 + 1.06} - 0.1 < 0.09,$$

y

$$|g_3(x_1, x_2, x_3)| = \frac{1}{20}e^{-x_1x_2} + \frac{10\pi - 3}{60} \leq \frac{1}{20}e + \frac{10\pi - 3}{60} < 0.61;$$

así que $-1 \leq g_i(x_1, x_2, x_3) \leq 1$, para toda $i = 1, 2, 3$. Por tanto, $\mathbf{G}(\mathbf{x}) \in D$ siempre que $\mathbf{x} \in D$.

Al obtener las cotas de las derivadas parciales de D, obtenemos

$$\left|\frac{\partial g_1}{\partial x_1}\right| = 0, \quad \left|\frac{\partial g_2}{\partial x_2}\right| = 0 \quad \text{y} \quad \left|\frac{\partial g_3}{\partial x_3}\right| = 0,$$

y también

$$\left|\frac{\partial g_1}{\partial x_2}\right| \leq \frac{1}{3}|x_3| \cdot |\operatorname{sen} x_2x_3| \leq \frac{1}{3}\operatorname{sen} = 1 < 0.281,$$

$$\left|\frac{\partial g_1}{\partial x_3}\right| \leq \frac{1}{3}|x_2| \cdot |\operatorname{sen} x_2x_3| \leq \frac{1}{3}\operatorname{sen} 1 < 0.281,$$

$$\left|\frac{\partial g_2}{\partial x_1}\right| = \frac{|x_1|}{9\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06}} < \frac{1}{9\sqrt{0.218}} < 0.238,$$

$$\left|\frac{\partial g_2}{\partial x_3}\right| = \frac{|\cos x_3|}{18\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06}} < \frac{1}{18\sqrt{0.218}} < 0.119,$$

$$\left|\frac{\partial g_3}{\partial x_1}\right| = \frac{|x_2|}{20}e^{-x_1x_2} \leq \frac{1}{20}e < 0.14,$$

y

$$\left|\frac{\partial g_3}{\partial x_2}\right| = \frac{|x_1|}{20}e^{-x_1x_2} \leq \frac{1}{20}e < 0.14.$$

Como las derivadas parciales de g_1, g_2 y g_3 están acotadas en D, el teorema 10.4 implica que estas funciones son continuas en D. En consecuencia, **G** es continua en D. Además, para todo $\mathbf{x} \in D$,

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right| \leq 0.281, \quad \text{para cada } i = 1, 2, 3 \quad \text{y} \quad j = 1, 2, 3,$$

y la condición de la segunda parte del teorema 10.6 se cumple con $K = 3(0.281) = 0.843$.

Del mismo modo se puede demostrar también que $\partial g_i/\partial x_j$ es continua en D para cada $i = 1, 2, 3$ y $j = 1, 2, 3$. (Esto se considera en el ejercicio 3.) En consecuencia, **G** tiene un único punto fijo en D y el sistema no lineal tiene una solución en D.

Nótese que el hecho de que **G** tenga una solución única en D no significa que la solución del sistema original sea única en este dominio, ya que la solución de x_2 en (10.4) requiere la elección de la raíz cuadrada principal. En el ejercicio 7(d) se analiza la situación que ocurre si, en cambio, en este paso se selecciona la raíz cuadrada negativa.

Para aproximar el punto fijo **p**, escogemos $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$. La sucesión de vectores generada por

$$x_1^{(k)} = \frac{1}{3}\cos x_2^{(k-1)} x_3^{(k-1)} + \frac{1}{6},$$

$$x_2^{(k)} = \frac{1}{9}\sqrt{\left(x_1^{(k-1)}\right)^2 + \operatorname{sen} x_3^{(k-1)} + 1.06} - 0.1$$

$$x_3^{(k)} = -\frac{1}{20}e^{-x_1^{(k-1)} x_2^{(k-1)}} - \frac{10\pi - 3}{60}$$

converge a la solución única de (10.4). En este ejemplo, la secuencia se generó hasta

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-5}.$$

Los resultados se incluyen en la tabla 10.1.

Tabla 10.1

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\lVert\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\rVert_\infty$
0	0.10000000	0.10000000	−0.10000000	
1	0.49998333	0.00944115	−0.52310127	0.423
2	0.49999593	0.00002557	−0.52336331	9.4×10^{-3}
3	0.50000000	0.00001234	−0.52359814	2.3×10^{-4}
4	0.50000000	0.00000003	−0.52359847	1.2×10^{-5}
5	0.50000000	0.00000002	−0.52359877	3.1×10^{-7}

Al utilizar la cota de error (10.3) con $K = 0.843$ obtenemos

$$\|\mathbf{x}^{(5)} - \mathbf{p}\|_\infty \leq \frac{(0.843)^5}{1 - 0.843}(0.423) < 1.15,$$

que no indica la exactitud real de $\mathbf{x}^{(5)}$ debido a la aproximación inicial inexacta. La solución real es

$$\mathbf{p} = \left(0.5, 0, -\frac{\pi}{6}\right)^t \approx (0.5, 0, -0.5235987757)^t,$$

y, por tanto, el error verdadero es

$$\|\mathbf{x}^{(5)} - \mathbf{p}\|_\infty \leq 2 \times 10^{-8}.$$

Una forma de acelerar la convergencia de la iteración de punto fijo consiste en usar las estimaciones más recientes de $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$ en vez de $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$ para calcular $x_i^{(k)}$, igual que en el método de Gauss-Seidel para los sistemas lineales. Entonces las ecuaciones componentes se transforman en

$$x_1^{(k)} = \frac{1}{3}\cos\left(x_2^{(k-1)} x_3^{(k-1)}\right) + \frac{1}{6},$$

$$x_2^{(k)} = \frac{1}{9}\sqrt{\left(x_1^{(k)}\right)^2 + \operatorname{sen} x_3^{(k-1)} + 1.06} - 0.1,$$

$$x_3^{(k)} = -\frac{1}{20}e^{-x_1^{(k)} x_2^{(k)}} - \frac{10\pi - 3}{60}.$$

Con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$, los resultados de estos cálculos se dan en la tabla 10.2.

La iteración $\mathbf{x}^{(4)}$ es exacta con una precisión de 10^{-7} en la norma l_∞; por tanto, en este problema la convergencia efectivamente se aceleró aplicando el método de Gauss-Seidel. No obstante, conviene recalcar que esa técnica no *siempre* la acelera.

Tabla 10.2

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\lVert\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\rVert_\infty$
0	0.10000000	0.10000000	−0.10000000	
1	0.49998333	0.02222979	−0.52304613	0.423
2	0.49997747	0.00002815	−0.52359807	2.2×10^{-2}
3	0.50000000	0.00000004	−0.52359877	2.8×10^{-5}
4	0.50000000	0.00000000	−0.52359877	3.8×10^{-8}

Maple ofrece la función `fsolve` para resolver sistemas de ecuaciones. El problema de punto fijo del ejemplo 2 puede resolverse con los siguientes comandos:

```
>g1:=x1=(2*cos(x2*x3)+1)/6;
>g2:=x2=sqrt(x^2+sen(x3)+1.06)/9-0.1;
>g3:=x3=-(3*exp(-x1*x2)+10*Pi-3)/60;
>fsolve({g1,g2,g3},{x1,x2,x3},{x1=-1..1,x2=-1..1,x3=-1..1});
```

Los tres primeros comandos definen el sistema y el último contiene el procedimiento `fsolve`. La respuesta mostrada en la pantalla es

$$\{x3 = -.5235987758, x1 = .5000000000, x2 = -.2102454409\ 10^{-10}\}$$

En general, `fsolve(eqns,vars,options)` resuelve el sistema de ecuaciones representadas por el parámetro `eqns`, para las variables representadas por el parámetro `vars`, bajo los parámetros opcionales representados por `options`. Con `options` especificamos una región donde se requiere la rutina para buscar una solución. Esta especificación no es obligatoria, y Maple determina su propio espacio de búsqueda, cuando se omiten las opciones.

CONJUNTO DE EJERCICIOS 10.1

1. Demuestre que la función $\mathbf{F}: \mathbb{R}^3 \to \mathbb{R}^3$ definida por

$$\mathbf{F}(x_1, x_2, x_3) = (x_1 + 2x_3, x_1 \cos x_2, x_2^2 + x_3)^t$$

es continua en cada punto de $\mathbb{R}^3$.

2. Dé un ejemplo de una función $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2$ que sea continua en cada punto de $\mathbb{R}^2$ excepto en $(1, 0)$.

3. Demuestre que las primeras derivadas parciales del ejemplo 2 son continuas en D.

4. El sistema no lineal

$$-x_1(x_1 + 1) + 2x_2 = 18,$$

$$(x_1 - 1) + (x_2 - 6)^2 = 25$$

tiene dos soluciones.

a. Aproxime gráficamente las soluciones.

b. Utilice las aproximaciones de la parte (a) como aproximaciones iniciales en una iteración funcional apropiada y determine las soluciones dentro de 10^{-5} en la norma l_∞.

5. El sistema no lineal

$$x_1^2 - 10x_1 + x_2^2 + 8 = 0,$$

$$x_1x_2^2 + x_1 - 10x_2 + 8 = 0$$

puede transformarse en el problema de punto fijo

$$x_1 = g_1(x_1, x_2) = \frac{x_1^2 + x_2^2 + 8}{10},$$

$$x_2 = g_1(x_1, x_2) = \frac{x_1x_2^2 + x_1 + 8}{10}.$$

a. Use el teorema 10.6 para demostrar que $\mathbf{G} = (g_1, g_2)^t$ que mapea $D \subset \mathbb{R}^2$ en $\mathbb{R}^2$ tiene un punto fijo en

$$D = \{(x_1, x_2)^t \mid 0 \le x_1, x_2 \le 1.5\}.$$

b. Aplique la iteración funcional para aproximar la solución.

c. ¿Acelera la convergencia el método de Gauss-Seidel?

6. El sistema no lineal

$$5\,x_1^2 - x_2^2 = 0$$

$$x_2 - 0.25(\operatorname{sen} x_1 + \cos x_2) = 0$$

tiene una solución cercana a $\left(\frac{1}{4}, \frac{1}{4}\right)^t$.

a. Encuentre una función $\mathbf{G}$ y un conjunto D en $\mathbb{R}^2$ de modo que $\mathbf{G}: D \to \mathbb{R}^2$ y $\mathbf{G}$ tenga un punto fijo único en D.

b. Aplique la iteración funcional para aproximar la solución con una exactitud de 10^{-5} en la norma l_∞.

c. ¿Acelera la convergencia el método de Gauss-Seidel?

7. Use el teorema 10.6 para demostrar que $\mathbf{G}: D \subset \mathbb{R}^3 \to \mathbb{R}^3$ tiene un punto fijo en D. Aplique la iteración funcional para aproximar la solución con una exactitud de 10^{-5} empleando $\|\cdot\|_\infty$.

   a. $$\mathbf{G}(x_1, x_2, x_3) = \left(\frac{\cos(x_2x_3) + 0.5}{3}, \frac{1}{25}\sqrt{x_1^2 + 0.3125} - 0.03, -\frac{1}{20}e^{-x_1x_2} - \frac{10\pi - 3}{60}\right)^t;$$
   $$D = \{(x_1, x_2, x_3)^t \mid -1 \le x_i \le 1, i = 1, 2, 3\}$$

   b. $$\mathbf{G}(x_1, x_2, x_3) = \left(\frac{13 - x_2^2 + 4x_3}{15}, \frac{11 + x_3 - x_1^2}{10}, \frac{22 + x_2^3}{25}\right)^t;$$
   $$D = \{(x_1, x_2, x_3)^t \mid 0 \le x_i \le 1.5, i = 1, 2, 3\}$$

   c. $$\mathbf{G}(x_1, x_2, x_3) = (1 - \cos(x_1x_2x_3), 1 - (1 - x_1)^{1/4} - 0.05x_3^2 + 0.15x_3, x_1^2 + 0.1x_2^2 - 0.01x_2 + 1)^t;$$
   $$D = \{(x_1, x_2, x_3)^t \mid -0.1 \le x_1 \le 0.1, -0.1 \le x_2 \le 0.3, 0.5 \le x_3 \le 1.1\}$$

   d. $$\mathbf{G}(x_1, x_2, x_3) = \left(\frac{1}{3}\cos(x_2 x_3) + \frac{1}{6}, -\frac{1}{9}\sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} - 0.1, -\frac{1}{20}e^{-x_1x_2} - \frac{10\pi - 3}{60}\right)^t;$$
   $$D = \{(x_1, x_2, x_3)^t \mid -1 \le x_i \le 1, i = 1, 2, 3\}$$

8. Aplique el método de Gauss-Seidel para aproximar los puntos fijos del ejercicio 7 con una exactitud de 10^{-5} empleando $\|\cdot\|_\infty$.

9. Use la iteración funcional para obtener soluciones a los siguientes sistemas no lineales, con una exactitud de 10^{-5}, empleando $\|\cdot\|_\infty$.

   a. $$x_1^2 + x_2^2 - x_1 = 0,$$
   $$x_1^2 - x_2^2 - x_2 = 0.$$

   b. $$3x_1^2 - x_2^2 = 0,$$
   $$3x_1x_2^2 - x_1^3 - 1 = 0.$$

   c. $$x_1^2 + x_2 - 37 = 0,$$
   $$x_1 - x_2^2 - 5 = 0,$$
   $$x_1 + x_2 + x_3 - 3 = 0$$

   d. $$x_1^2 + 2x_2^2 - x_2 - 2x_3 = 0,$$
   $$x_1^2 - 8x_2^2 + 10x_3 = 0,$$
   $$\frac{x_1^2}{7x_2x_3} - 1 = 0.$$

10. Demuestre que la función $\mathbf{F}$ que mapea $D \subset \mathbb{R}^n$ en $\mathbb{R}^n$ es continua en $\mathbf{x}_0 \in D$, exactamente cuando, dado un número cualquiera $\varepsilon > 0$, se pueda encontrar un número $\delta > 0$ con la propiedad de que para cualquier norma vectorial $\|\cdot\|$,
$$\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)\| < \varepsilon,$$
siempre que $\mathbf{x} \in D$ y $\|\mathbf{x} - \mathbf{x}_0\| < \delta$.

11. En el ejercicio 8 de la sección 5.9 vimos el problema de predecir la población de dos especies que compiten por el mismo suministro de comida. Supusimos, entonces, que las poblaciones podían predecirse resolviendo el sistema de ecuaciones
$$\frac{dx_1(t)}{dt} = x_1(t)(4 - 0.0003x_1(t) - 0.0004x_2(t))$$

y

$$\frac{dx_2(t)}{dt} = x_2(t)(2 - 0.0002x_1(t) - 0.0001x_2(t)).$$

En este ejercicio nos gustaría considerar el problema de determinar las poblaciones de equilibrio de las dos especies. El criterio matemático que se debe satisfacer para que las poblaciones estén en equilibrio es que, simultáneamente,

$$\frac{dx_1(t)}{dt} = 0 \quad \text{y} \quad \frac{dx_2(t)}{dt} = 0.$$

Esto sucede cuando la primera especie está extinta y la segunda tiene una población de 20 000, o cuando la segunda está extinta y la primera tiene una población de 13 333. ¿Puede haber equilibrio en cualquier otra situación?

10.2 Método de Newton

El problema del ejemplo 2 de la sección anterior se convierte en un problema de punto fijo convergente, si se resuelven algebraicamente las tres ecuaciones para las tres variables x_1, x_2 y x_3. Sin embargo, esta técnica rara vez tiene éxito. En esta sección estudiaremos un procedimiento algorítmico para efectuar la transformación en una situación más general.

Para construir el algoritmo que nos lleva a un método de punto fijo apropiado en el caso unidimensional, obtuvimos una función ϕ con la propiedad de que

$$g(x) = x - \phi(x)f(x)$$

da una convergencia cuadrática en el punto fijo p de la función g (véase la sección 2.4). A partir de esta condición, el método de Newton evoluciona al seleccionar $\phi(x) = 1/f'(x)$, suponiendo que $f'(x) \neq 0$.

La aplicación de un procedimiento semejante en el caso n-dimensional incluye una matriz

$$A(\mathbf{x}) = \begin{bmatrix} a_{11}(\mathbf{x}) & a_{12}(\mathbf{x}) & \dots & a_{1n}(\mathbf{x}) \\ a_{21}(\mathbf{x}) & a_{22}(\mathbf{x}) & \dots & a_{2n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ a_{n1}(\mathbf{x}) & a_{n2}(\mathbf{x}) & \dots & a_{nn}(\mathbf{x}) \end{bmatrix} \tag{10.5}$$

donde todos los elementos $a_{ij}(\mathbf{x})$ son una función de $\mathbb{R}^n$ en $\mathbb{R}$. Esto requiere obtener $A(\mathbf{x})$ de modo que

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x})$$

dé la convergencia cuadrática a la solución de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, suponiendo que $A(\mathbf{x})$ es no singular en el punto fijo $\mathbf{p}$ de $\mathbf{G}$.

El siguiente teorema es paralelo al teorema 2.8 de la sección 2.4. Su demostración requiere la capacidad de expresar $\mathbf{G}$ en términos de su serie de Taylor en n variables alrededor de $\mathbf{p}$.

Teorema 10.7 Supongamos que $\mathbf{p}$ es una solución de $\mathbf{G}(\mathbf{x}) = \mathbf{x}$. Si existe un número $\delta > 0$ con la propiedad de que

(i) $\partial g_i/\partial x_j$ sea continua en $N_\delta = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{p}\| < \delta\}$ para toda $i = 1, 2, \ldots, n$ y toda $j = 1, 2, \ldots, n$;

(ii) $\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)$ sea continua y $|\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)| \le M$ para alguna constante M siempre que $\mathbf{x} \in N_\delta$, para toda $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, y toda $k = 1, 2, \ldots, n$;

(iii) $\partial g_i(\mathbf{p})/\partial x_k = 0$ para toda $i = 1, 2, \ldots, n$ y toda $k = 1, 2, \ldots, n$.

Entonces existe un número $\hat{\delta} \le \delta$ tal que la sucesión generada por $\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)})$ converge cuadráticamente a $\mathbf{p}$ para cualquier elección de $\mathbf{x}^{(0)}$ a condición de que $\|\mathbf{x}^{(0)} - \mathbf{p}\| < \hat{\delta}$. Más aún,

$$\|\mathbf{x}^{(k)} - \mathbf{p}\|_\infty \le \frac{n^2 M}{2} \|\mathbf{x}^{(k-1)} - \mathbf{p}\|_\infty^2, \quad \text{para toda } k \ge 1. \qquad \blacksquare$$

Para utilizar el teorema 10.7 supongamos que $A(\mathbf{x})$ es una matriz de $n \times n$ de funciones de $\mathbb{R}^n$ a $\mathbb{R}$ en la forma de la ecuación (10.5), cuyos elementos específicos se escogerán más adelante. Supongamos además que $A(\mathbf{x})$ es no singular cerca de una solución $\mathbf{p}$ de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, y denotemos con $b_{ij}(\mathbf{x})$ el elemento de $A(\mathbf{x})^{-1}$ en el i-ésimo renglón y en la j-ésima columna.

Dado que $\mathbf{G}(\mathbf{x}) = \mathbf{x} - A(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x})$, tenemos $g_i(\mathbf{x}) = x_i - \sum_{j=1}^{n} b_{ij}(\mathbf{x}) f_j(\mathbf{x})$ y

$$\frac{\partial g_i}{\partial x_k}(\mathbf{x}) = \begin{cases} 1 - \displaystyle\sum_{j=1}^{n}\left(b_{ij}(\mathbf{x})\frac{\partial f_j}{\partial x_k}(\mathbf{x}) + \frac{\partial b_{ij}}{\partial x_k}(\mathbf{x}) f_j(\mathbf{x})\right), & \text{si } i = k, \\ -\displaystyle\sum_{j=1}^{n}\left(b_{ij}(\mathbf{x})\frac{\partial f_j}{\partial x_k}(\mathbf{x}) + \frac{\partial b_{ij}}{\partial x_k}(\mathbf{x}) f_j(\mathbf{x})\right), & \text{si } i \ne k. \end{cases}$$

El teorema 10.7 implica que necesitamos $\partial g_i(\mathbf{p})/\partial x_k = 0$ para toda $i = 1, 2, \ldots, n$ y toda $k = 1, 2, \ldots, n$. Esto significa que, para toda $i = k$,

$$0 = 1 - \sum_{j=1}^{n} b_{ij}(\mathbf{p}) \frac{\partial f_j}{\partial x_i}(\mathbf{p}),$$

por lo que

$$\sum_{j=1}^{n} b_{ij\,614}(\mathbf{p}) \frac{\partial f_j}{\partial x_i}(\mathbf{p}) = 1. \tag{10.6}$$

Cuando $k \ne i$,

$$0 = -\sum_{j=1}^{n} b_{ij}(\mathbf{p}) \frac{\partial f_j}{\partial x_k}(\mathbf{p}),$$

por lo que

$$\sum_{j=1}^{n} b_{ij}(\mathbf{p}) \frac{\partial f_j}{\partial x_k}(\mathbf{p}) = 0. \tag{10.7}$$

Al definir la matriz $J(\mathbf{x})$ por medio de

$$J(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}) \end{bmatrix}, \tag{10.8}$$

vemos que las condiciones (10.6) y (10.9) requieren

$$A(\mathbf{p})^{-1}J(\mathbf{p}) = I, \quad \text{la matriz identidad,}$$

por lo que

$$A(\mathbf{p}) = J(\mathbf{p}).$$

En consecuencia, una elección apropiada de $A(\mathbf{x})$ es $A(\mathbf{x}) = J(\mathbf{x})$, dado que entonces se cumple la condición (iii) del teorema 10.7.

La función $\mathbf{G}$ está definida por

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - J(\mathbf{x})^{-1}\mathbf{F}(\mathbf{x}),$$

y el procedimiento de la iteración funcional pasa de seleccionar $\mathbf{x}^{(0)}$ a generar, para $k \geq 1$,

$$\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)}) = \mathbf{x}^{(k-1)} - J(\mathbf{x}^{(k-1)})^{-1}\mathbf{F}(\mathbf{x}^{(k-1)}). \tag{10.9}$$

A esto se le llama **método de Newton para sistemas no lineales,** y generalmente se espera que dé una convergencia cuadrática, siempre y cuando se conozca un valor inicial suficientemente preciso y exista $J(\mathbf{p})^{-1}$.

A la matriz $J(\mathbf{x})$ se le llama **matriz Jacobiana** y tiene varias aplicaciones en el análisis. En particular, quizá el lector esté familiarizado con ella debido a su aplicación en la integración múltiple de una función de varias variables, en una región que requiere efectuar un cambio de variables.

La debilidad del método de Newton se debe a la necesidad de calcular e invertir la matriz $J(\mathbf{x})$ en cada paso. En la práctica, el cálculo explícito de $J(\mathbf{x})^{-1}$ se evita efectuando la operación en dos pasos. Primero, encontramos un vector $\mathbf{y}$ que satisfaga $J(\mathbf{x}^{(k-1)})\mathbf{y} = -F(\mathbf{x}^{(k-1)})$. Una vez hecho esto, obtenemos la nueva aproximación $\mathbf{x}^{(k)}$ agregando $\mathbf{y}$ a $\mathbf{x}^{(k-1)}$. En el algoritmo 10.1 se emplea este procedimiento de dos pasos.

ALGORITMO **10.1**

Método de Newton para sistemas

Para aproximar la solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de ecuaciones e incógnitas; aproximación inicial $\mathbf{x} = (x_1, \ldots, x_n)^t$,

tolerancia TOL; número máximo de iteraciones N.

SALIDA solución aproximada $\mathbf{x} = (x_1, \ldots, x_n)^t$ o un mensaje de que se rebasó el número de iteraciones.

Paso 1 Tome $k = 1$.

Paso 2 Mientras ($k \leq N$) haga los pasos 3-7.

Paso 3 Calcule $\mathbf{F}(\mathbf{x})$ y $J(\mathbf{x})$, donde $J(\mathbf{x})_{i,j} = (\partial f_i(\mathbf{x})/\partial x_j)$ para $1 \leq i, j \leq n$.

Paso 4 Resuelva el sistema lineal $n \times n$ $J(\mathbf{x})\mathbf{y} = -\mathbf{F}(\mathbf{x})$.

Paso 5 Tome $\mathbf{x} = \mathbf{x} + \mathbf{y}$.

Paso 6 Si $\|\mathbf{y}\| < TOL$, entonces SALIDA ($\mathbf{x}$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 7 Tome $k = k + 1$

Paso 8 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

EJEMPLO 1 En el ejemplo 2 de la sección 10.1 demostramos que el sistema no lineal

$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

tiene una solución aproximada en $(0.5, 0, -0.52359877)^t$. Aplicaremos el método de Newton para obtener esta aproximación cuando la aproximación inicial es $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$ y

$$\mathbf{F}(x_1, x_2, x_3) = (f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3), f_3(x_1, x_2, x_3))^t,$$

donde

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - \frac{1}{2},$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06,$$

y

$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3}.$$

La matriz jacobiana $J(\mathbf{x})$ para este sistema es

$$J(x_1, x_2, x_3) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_2 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}$$

y

$$\begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix} = \begin{bmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{bmatrix} + \begin{bmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{bmatrix},$$

donde

$$\begin{bmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{bmatrix} = -\left(J\left(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}\right)\right)^{-1} \mathbf{F}\left(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}\right).$$

Por tanto, en el k-ésimo paso, debemos resolver el sistema lineal $J(\mathbf{x}^{(k-1)}(\mathbf{y}^{(k-1)} = -\mathbf{F}(\mathbf{x}^{(k-1)})$ donde

$$(J(\mathbf{x}^{(k-1)}) = \begin{bmatrix} 3 & x_3^{(k-1)} \operatorname{sen} x_2^{(k-1)} x_3^{(k-1)} & x_2^{(k-1)} \operatorname{sen} x_2^{(k-1)} x_3^{(k-1)} \\ 2x_1^{(k-1)} & -162(x_2^{(k-1)} + 0.1) & \cos x_3^{(k-1)} \\ -x_2^{(k-1)} e^{-x_1^{(k-1)} x_2^{(k-1)}} & -x_1^{(k-1)} e^{-x_1^{(k-1)} x_2^{(k-1)}} & 20 \end{bmatrix},$$

$$\mathbf{y}^{(k-1)}) = \begin{bmatrix} y_1^{(k-1)} \\ y_2^{(k-1)} \\ y_3^{(k-1)} \end{bmatrix},$$

y

$$\mathbf{F}(\mathbf{x}^{(k-1)}) = \begin{bmatrix} 3x_1^{(k-1)} - \cos x_2^{(k-1)} x_3^{(k-1)} - \frac{1}{2} \\ \left(x_1^{(k-1)}\right)^2 - 81\left(x_2^{(k-1)} + 0.1\right)^2 + \operatorname{sen} x_3^{(k-1)} + 1.06 \\ e^{-x_1^{(k-1)} x_2^{(k-1)}} + 20x_3^{(k-1)} + \frac{10\pi - 3}{3} \end{bmatrix}.$$

En la tabla 10.3 se proporcionan los resultados que se obtienen al usar este procedimiento iterativo. ■

Tabla 10.3

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\lVert\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\rVert_\infty$
0	0.10000000	0.10000000	−0.10000000	
1	0.50003702	0.01946686	−0.52152047	0.422
2	0.50004593	0.00158859	−0.52355711	1.79×10^{-2}
3	0.50000034	0.00001244	−0.52359845	1.58×10^{-3}
4	0.50000000	0.00000000	−0.52359877	1.24×10^{-5}
5	0.50000000	0.00000000	−0.52359877	0

El ejemplo anterior ilustra el hecho de que el método de Newton puede converger muy rápidamente una vez conseguida una aproximación que esté cerca de la solución verdadera. Pero no siempre es fácil determinar los valores iniciales con que se llegará a una solución, y el método es relativamente costoso. En la siguiente sección estudiaremos un método que nos permite superar esta última debilidad. Por lo general pueden obtenerse buenos valores iniciales con el método que expondremos en la sección 10.4.

La aproximación inicial a las soluciones de sistemas no lineales de 2×2, y a menudo de 3×3, pueden obtenerse por medio de las características de graficación que ofrece Maple. El sistema no lineal

$$x_1^2 - x_2^2 + 2x_2 = 0,$$
$$2x_1 + x_2^2 - 6 = 0$$

tiene las dos soluciones $(0.625204094, 2.179355825)^t$ y $(2.109511920, -1.334532188)^t$. Si queremos utilizar Maple, antes debemos definir las dos ecuaciones

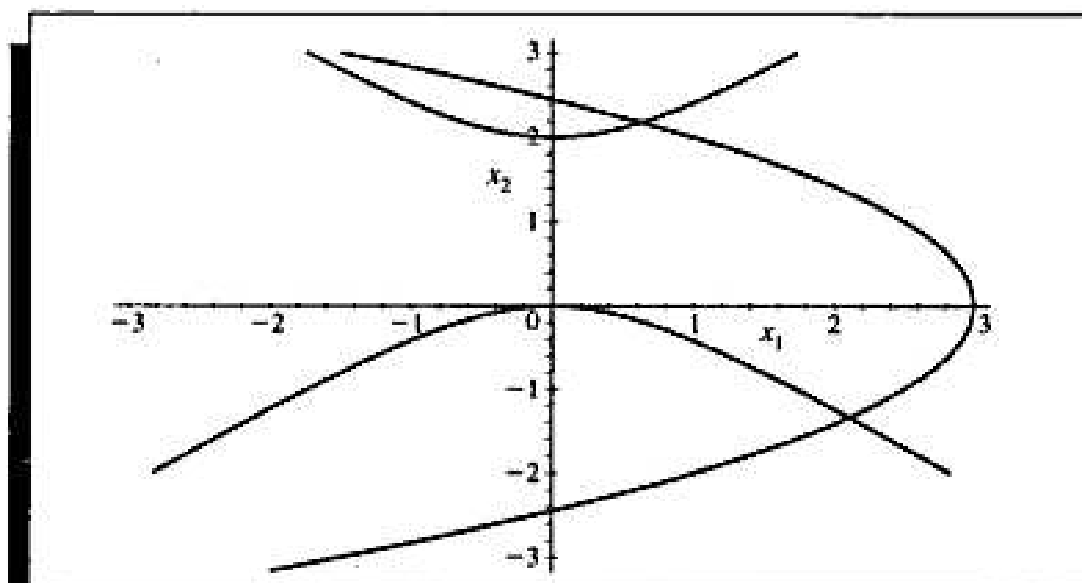
```
>eq1:=x1^2-x2^2+2*x2=0;
>eq2:=2*x1+x2^2-6=0;
```

Si queremos obtener una gráfica de las dos ecuaciones para $-3 \le x_1, x_2 \le 3$, introducimos los comandos

```
>with(plots):
>implicitplot({eq1,eq2},x1=-3..3,x2=-3..3);
```

La gráfica de la figura 10.2 nos permite estimar que hay soluciones cercanas a $(0.64, 2.2)^t$ y $(2.1, -1.3)^t$. Esto nos da buenos valores iniciales para el método de Newton.

FIGURA 10.2



El problema es más difícil en tres dimensiones. Considere el sistema no lineal

$$2x_1 - 3x_2 + x_3 - 4 = 0,$$
$$2x_1 + x_2 - x_3 + 4 = 0,$$
$$x_1^2 + x_2^2 + x_3^2 - 4 = 0.$$

Defina tres ecuaciones usando los comandos de Maple

```
>eq1:=2*x1-3*x2+x3-4=0;
>eq2:=2*x1+x2-x3+4=0;
>eq3:=x1^2+x2^2+x3^2-4=0;
```

La tercera ecuación describe una esfera de radio 2 y centro (0, 0, 0), así que `x1`, `x2` y `x3` están en $[-2, 2]$. Los comandos de Maple con que se produce la gráfica en este caso son

```
>with(plots);
>implicitplot3d({eq1,eq2,eq3},x1=-2..2,x2=-2..2,x3=-2..2);
```

Maple cuenta con varias opciones de graficación tridimensional para aislar una solución en el sistema no lineal. Por ejemplo, podemos girar la gráfica para visualizar mejor las secciones de las superficies. Después podemos hacer un acercamiento (*zoom*) a las regiones donde se hallan las intersecciones y modificar la forma de presentación de los ejes para conseguir una vista más exacta de las coordenadas de la intersección. Para este problema, una aproximación inicial razonable es $(x_1, x_2, x_3)^t = (-0.5, -1.5, 1.5)^t$.

CONJUNTO DE EJERCICIOS 10.2

1. Mediante el método de Newton con $\mathbf{x}^{(0)} = \mathbf{0}$ calcule $\mathbf{x}^{(2)}$ para los siguientes sistemas no lineales.

a.
$$4x_1^2 - 20x_1 + \frac{1}{4}x_2^2 + 8 = 0,$$
$$\frac{1}{2}x_1x_2^2 + 2x_1 - 5x_2 + 8 = 0.$$

b.
$$\operatorname{sen}(4\pi x_1x_2) - 2x_2 - x_1 = 0,$$
$$\left(\frac{4\pi - 1}{4\pi}\right)(e^{2x_1} - e) + 4ex_2^2 - 2ex_1 = 0.$$

c.
$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$
$$4x_1^2 - 625x_2^2 + 2x_2 - 1 = 0,$$
$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

d.
$$x_1^2 + x_2 - 37 = 0,$$
$$x_1 - x_2^2 - 5 = 0,$$
$$x_1 + x_2 + x_3 - 3 = 0.$$

2. Use las opciones de graficación de Maple para aproximar las soluciones de los siguientes problemas no lineales.

a.
$$x_1(1 - x_1) + 4x_2 = 12,$$
$$(x_1 - 2)^2 + (2x_2 - 3)^2 = 25.$$

b.
$$5x_1^2 - x_2^2 = 0,$$
$$x_2 - 0.25(\operatorname{sen} x_1 + \cos x_2) = 0.$$

c.
$$15x_1 + x_2^2 - 4x_3 = 13,$$
$$x_1^2 + 10x_2 - x_3 = 11,$$
$$x_2^3 - 25x_3 = -22.$$

d.
$$10x_1 - 2x_2^2 + x_2 - 2x_3 - 5 = 0,$$
$$8x_2^2 + 4x_3^2 - 9 = 0,$$
$$8x_2x_3 + 4 = 0.$$

3. Use el método de Newton para encontrar una solución a los siguientes sistemas no lineales con la aproximación inicial dada. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-6}$.

a.
$$3x_1^2 - x_2^2 = 0,$$
$$3x_1x_2^2 - x_1^3 - 1 = 0.$$
Use $\mathbf{x}^{(0)} = (1, 1)^t$.

b.
$$\ln(x_1^2 + x_2^2) - \operatorname{sen}(x_1x_2) = \ln 2 + \ln \pi,$$
$$e^{x_1 - x_2} + \cos(x_1x_2) = 0.$$
Use $\mathbf{x}^{(0)} = (2, 2)^t$.

c.
$$x_1^3 + x_1^2x_2 - x_1x_3 + 6 = 0,$$
$$e^{x_1} + e^{x_2} - x_3 = 0,$$
$$x_2^2 - 2x_1x_3 = 4.$$
Use $\mathbf{x}^{(0)} = (-1, -2, 1)^t$.

d.
$$6x_1 - 2\cos(x_2x_3) - 1 = 0,$$
$$9x_2 + \sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} + 0.9 = 0,$$
$$60x_3 + 3e^{-x_1x_2} + 10\pi - 3 = 0.$$
Use $\mathbf{x}^{(0)} = (0, 0, 0)^t$.

4. Use las respuestas obtenidas en el ejercicio 2 como aproximaciones iniciales al método de Newton. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-6}$.

5. El sistema no lineal

$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 625x_2^2 - \frac{1}{4} = 0,$$

$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

tiene una matriz jacobiana singular en la solución. Aplique el método de Newton con $\mathbf{x}^{(0)} = (1, 1 -1)^t$. Observe que la convergencia puede ser lenta o no ocurrir después de realizar una cantidad razonable de iteraciones.

6. El sistema no lineal

$$4x_1 - x_2 + x_3 = x_1x_4,$$
$$-x_1 + 3x_2 - 2x_3 = x_2x_4,$$
$$x_1 - 2x_2 + 3x_3 = x_3x_4,$$
$$x_1^2 + x_2^2 + x_3^2 = 1$$

tiene seis soluciones.

a. Muestre que si $(x_1, x_2, x_3, x_4)^t$ es una solución, entonces $(-x_1, -x_2, -x_3, -x_4)^t$ también es una solución.

b. Use el método de Newton tres veces para aproximar todas las soluciones. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-5}$.

7. Muestre que cuando $n = 1$, el método de Newton dado por la ecuación (10.9) se reduce al método de Newton ya conocido dado por la ecuación (2.5).

8. ¿A qué se reduce el método de Newton para el sistema lineal $A\mathbf{x} = \mathbf{b}$ dado por

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

donde A es una matriz no singular?

9. C. Chiarella, W. Charlton y A. W. Roberts [CCR], al calcular la forma de un sumidero de descarga de flujo por gravedad que reducirá al máximo el tiempo de tránsito de las partículas granulares descargadas, resuelven la siguiente ecuación por medio del método de Newton:

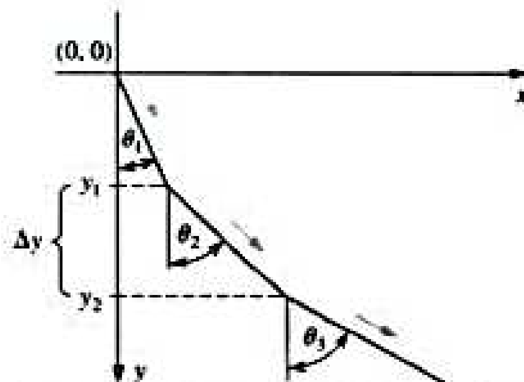
(i) $f_n(\theta_1, \ldots, \theta_N) = \dfrac{\text{sen}\ \theta_{n+1}}{v_{n+1}}(1 - \mu w_{n+1}) - \dfrac{\text{sen}\ \theta_n}{v_n}(1 - \mu w_n) = 0$, para toda $n = 1, 2, \ldots, N-1$.

(ii) $f_N(\theta_1, \ldots, \theta_N) = \Delta y \sum_{i=1}^{N} \tan \theta_i - X = 0$, donde

a. $v_n^2 = v_0^2 - 2gn\Delta y - 2\mu\Delta y \sum_{j=1}^{n} \dfrac{1}{\cos \theta_j}$, para toda $n = 1, 2, \ldots, N$, y

b. $w_n = -\Delta y v_n \sum_{i=1}^{N} \dfrac{1}{v_i^3 \cos \theta_i}$, para toda $n = 1, 2, \ldots, N$.

La constante v_0 es la velocidad inicial del material granular, X es la coordenada x del extremo final del sumidero, μ es la fuerza de fricción, N es el número de segmentos del sumidero y g es la constante gravitacional 32.17 pie/s^2. La variable θ_i es el ángulo del i-ésimo segmento de sumidero respecto a la vertical, como se aprecia en la figura siguiente, y v_i es la velocidad de la partícula en el i-ésimo segmento del sumidero. Resuelva (i) y (ii) para $\theta = (\theta_1, \ldots, \theta_N)^t$ con $\mu = 0$, $X = 2$, $\Delta y = 0.2$, $N = 20$ y $v_0 = 0$, donde los valores de v_n y w_n pueden obtenerse directamente de (a) y (b). Itere hasta que $\|\theta^{(k)} - \theta^{(k-1)}\|_\infty < 10^{-2}$.



10. La presión requerida para enterrar un objeto grande y pesado en un suelo blando homogéneo que se encuentra sobre una base de suelo duro puede predecirse a partir de la presión necesaria para enterrar objetos más pequeños en el mismo terreno. En concreto, la presión p requerida para enterrar una placa circular de radio r a una distancia d en el suelo blando, donde la base dura se encuentra a una distancia $D > d$ debajo de la superficie, puede aproximarse mediante una ecuación de la forma

$$p = k_1 e^{k_2 r} + k_3 r,$$

donde k_1, k_2 y k_3 son constantes, con $k_2 > 0$ que depende de d y de la consistencia del terreno pero no del radio de la placa. (Véase [Bek, pp. 89–94].)

a. Calcule los valores de k_1, k_2 y k_3 si suponemos que una placa cuyo radio es de 1 plg requiere una presión de 10 lb/plg^2 para enterrarse 1 pie en un campo fangoso, una placa cuyo radio es de 2 plg requiere una presión de 12 lb/plg^2 para enterrarse 1 pie y una placa de 3 plg de radio requiere una presión de 15 lb/plg^2 para enterrarse esta distancia (suponiendo que el lodo tiene una profundidad de más de 1 pie).

b. Use los cálculos de la parte (a) para predecir el tamaño mínimo de la placa circular que se necesitará para sostener una carga de 500 lb en este campo, con un hundimiento menor a 1 pie.

11. Un interesante experimento biológico (véase [Schr2]) es la determinación de la temperatura máxima del agua, X_M, en la que varias especies de hidra pueden sobrevivir sin que su esperanza de vida disminuya. Una forma de resolver este problema consiste en aplicar un ajuste ponderado de mínimos cuadrados de la forma $f(x) = y = a/(x - b)^c$ a un conjunto de datos experimentales. Los valores x de los datos se refieren a la temperatura del agua. La constante b es la asíntota de la gráfica de f y, por tanto, es una aproximación a X_M.

a. Demuestre que la elección de a, b y c para disminuir al mínimo

$$\sum_{i=1}^{n}\left[w_i y_i - \frac{a}{(x_i - b)^c}\right]^2$$

se reduce a resolver el sistema no lineal

$$a = \sum_{i=1}^{n} \frac{w_i y_i}{(x_i - b)^c} \Bigg/ \sum_{i=1}^{n} \frac{1}{(x_i - b)^{2c}},$$

$$0 = \sum_{i=1}^{n} \frac{w_i y_i}{(x_i - b)^c} \cdot \sum_{i=1}^{n} \frac{1}{(x_i - b)^{2c+1}} - \sum_{i=1}^{n} \frac{w_i y_i}{(x_i - b)^{c+1}} \cdot \sum_{i=1}^{n} \frac{1}{(x_i - b)^{2c}},$$

$$0 = \sum_{i=1}^{n} \frac{w_i y_i}{(x_i - b)^c} \cdot \sum_{i=1}^{n} \frac{\ln (x_i - b)}{(x_i - b)^{2c}} - \sum_{i=1}^{n} \frac{w_i y_i \ln(x_i - b)}{(x_i - b)^c} \cdot \sum_{i=1}^{n} \frac{1}{(x_i - b)^{2c}}.$$

b. Con los siguientes datos, resuelva el sistema no lineal para las especies. Utilice los pesos $w_i = \ln y_i$.

i	1	2	3	4
y_i	2.40	3.80	4.75	21.60
x_i	31.8	31.5	31.2	30.2

10.3 Métodos cuasi-Newton

Un punto débil importante del método de Newton para resolver sistemas de ecuaciones no lineales es el requisito de que, en cada iteración, es necesario calcular una matriz jacobiana y resolver un sistema lineal de $n \times n$ que la contiene. Para ejemplificar la importancia de tal debilidad, consideremos los cálculos que trae consigo una iteración de dicho método. La matriz jacobiana asociada a un sistema de n ecuaciones no lineales escritas en la forma $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ requiere determinar y evaluar las n^2 derivadas parciales de las n funciones componentes de $\mathbf{F}$. En casi todos los casos resulta incómodo evaluar exactamente las derivadas parciales, aunque el problema se facilita ahora con el empleo generalizado de los sistemas de cómputo simbólicos, como Maple.

Cuando no es práctico efectuar la evaluación exacta, podemos usar las aproximaciones de diferencia finita a las derivadas parciales. Por ejemplo,

$$\frac{\partial f_j}{\partial x_k}(\mathbf{x}^{(i)}) \approx \frac{f_j(\mathbf{x}^{(i)} + \mathbf{e}_k h) - f_j(\mathbf{x}^{(i)})}{h}, \tag{10.10}$$

donde h es pequeña en valor absoluto y $\mathbf{e}_k$ es el vector cuyo único elemento distinto de cero es un 1 de la k-ésima coordenada. Sin embargo, esta aproximación requiere efectuar, al menos, n^2 evaluaciones de funciones escalares para aproximar la matriz jacobiana y no disminuye la cantidad de cálculos, casi siempre es necesario $O(n^3)$ para resolver el sistema lineal que contiene esta matriz jacobiana aproximada. Por consiguiente, el total de cálculos que se requiere para una sola iteración del método de Newton es al menos $n^2 + n$ evaluaciones de funciones escalares (n^2 para evaluar la matriz jacobiana y n para evaluar $\mathbf{F}$), junto con $O(n^3)$ operaciones aritméticas para resolver el sistema lineal. Esta cantidad de cálculos es muy grande, excepto en el caso de los valores relativamente pequeños de n y de funciones escalares fáciles de evaluar.

En esta sección estudiaremos una generalización del método de la secante a los sistemas de ecuaciones no lineales, técnica denominada **método de Broyden** (véase [Broy]). El método requiere sólo n evaluaciones de funciones escalares por iteración y también disminuye el número de cálculos aritméticos a $O(n^2)$. Pertenece a una clase de técnicas denominadas *actualizaciones de secante con cambio mínimo* que dan origen a los algoritmos llamados **cuasi-Newton**. Estos métodos reemplazan a la matriz jacobiana en el método de Newton con una matriz de aproximación que se actualiza en cada iteración. Su desventaja radica en que se pierde la convergencia cuadrática de Newton, al ser sustituida por una convergencia denominada **superlineal**, la cual implica que

$$\lim_{i\to\infty} \frac{\|\mathbf{x}^{(i+1)} - \mathbf{p}\|}{\|\mathbf{x}^{(i)} - \mathbf{p}\|} = 0,$$

donde $\mathbf{p}$ denota la solución de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ y $\mathbf{x}^{(i)}$ y $\mathbf{x}^{(i+1)}$ son aproximaciones consecutivas de $\mathbf{p}$.

En la mayoría de las aplicaciones, la reducción a la convergencia superlineal es una manera más que aceptable para disminuir los cálculos. Otra desventaja más de los métodos cuasi-Newton es que, a diferencia del método de Newton, no se corrigen a sí mismos. El método de Newton generalmente corregirá el error del redondeo con iteraciones sucesivas, no así el método de Broyden, salvo que se incorporen medidas especiales de corrección.

Para describir el método de Broyden, supongamos que se da una aproximación inicial $\mathbf{x}^{(0)}$ a la solución $\mathbf{p}$ de $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Calculamos la siguiente aproximación $\mathbf{x}^{(1)}$ como lo hacemos en el método de Newton o, si es difícil determinar exactamente $J(\mathbf{x}^{(0)})$, utilizaremos las ecuaciones de diferencia dadas por (10.10) para aproximar las derivadas parciales. No obstante, para calcular $\mathbf{x}^{(2)}$ nos apartamos del método de Newton y examinamos el método de la secante para una sola ecuación no lineal. En el método de la secante utilizamos la aproximación

$$f'(x_1) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

como sustituto de $f'(x_1)$ en el método de Newton. En el caso de los sistemas no lineales, $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ es un vector, y el cociente correspondiente está indefinido. Pero el método procede de manera semejante al método de Newton, porque reemplazamos la matriz $J(\mathbf{x}^{(1)})$ por una matriz A_1 que tiene la propiedad de que

$$A_1(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}) = \mathbf{F}(\mathbf{x}^{(1)}) - \mathbf{F}(\mathbf{x}^{(0)}). \tag{10.11}$$

Todo vector distinto de cero de $\mathbf{R}^n$ puede escribirse como la suma de un múltiplo de $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ y de un múltiplo de un vector en el complemento ortogonal de $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$. Por tanto, si queremos definir específicamente a la matriz A_1, necesitamos determinar cómo actúa sobre el complemento ortogonal de $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$. Dado que no se tiene información sobre el cambio de $\mathbf{F}$ en una dirección ortogonal con $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$, requerimos que

$$A_1\mathbf{z} = J(\mathbf{x}^{(0)})\mathbf{z}, \quad \text{siempre que } (\mathbf{x}^{(1)} - \mathbf{x}^{(0)})^t\mathbf{z} = 0. \tag{10.12}$$

Esta condición especifica que ningún vector ortogonal a $\mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ se ve afectado por la actualización de $J(\mathbf{x}^{(0)})$, que sirvió para calcular $\mathbf{x}^{(1)}$, a A_1, con que se determinó $\mathbf{x}^{(2)}$.

Las condiciones (10.11) y (10.12) definen de manera única a A_1 (véase a [DM]) como

$$A_1 = J(\mathbf{x}^{(0)}) + \frac{[\mathbf{F}(\mathbf{x}^{(1)}) - \mathbf{F}(\mathbf{x}^{(0)}) - J(\mathbf{x}^{(0)})(\mathbf{x}^{(1)} - \mathbf{x}^{(0)})](\mathbf{x}^{(1)} - \mathbf{x}^{(0)})^t}{\|(\mathbf{x}^{(1)} - \mathbf{x}^{(0)})\|_2^2}.$$

Esta matriz es la que se usa en lugar de $J(\mathbf{x}^{(1)})$ para determinar $\mathbf{x}^{(2)}$ como

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - A_1^{-1}\mathbf{F}(\mathbf{x}^{(1)}).$$

Una vez determinada $\mathbf{x}^{(2)}$, el método se repite hasta determinar $\mathbf{x}^{(3)}$, usando A_1 en lugar de $A_0 \equiv J(\mathbf{x}^{(0)})$ y con $\mathbf{x}^{(2)}$ y $\mathbf{x}^{(1)}$ en lugar de $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(0)}$. En general, una vez que hemos determinado $\mathbf{x}^{(i)}$, calculamos $\mathbf{x}^{(i+1)}$ por medio de

$$A_i = A_{i-1} + \frac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{\|\mathbf{s}_i\|_2^2} \tag{10.13}$$

y

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - A_i^{-1}\,\mathbf{F}(\mathbf{x}^{(i)}), \tag{10.14}$$

donde la notación $y_i = \mathbf{F}(\mathbf{x}^{(i)}) - \mathbf{F}(\mathbf{x}^{(i+1)})$ y $\mathbf{s}_i = \mathbf{x}^{(i)} - \mathbf{x}^{(i+1)}$ se introduce en (10.13) para simplificar las ecuaciones.

Si el método se aplica como se describe en las ecuaciones (10.13) y (10.14), el número de evaluaciones de funciones escalares disminuye de $n^2 + n$ a n (las necesarias para evaluar $\mathbf{F}(\mathbf{x}^{(i)})$), pero todavía se requieren $O(n^3)$ cálculos para resolver el sistema lineal asociado de $n \times n$ (véase paso 4 del algoritmo 10.1).

$$A_i\,\mathbf{s}_{i+1} = -\mathbf{F}(x^{(i)}). \tag{10.15}$$

El uso del método en esta forma no se justificaría, debido a la reducción a la convergencia superlineal a partir de la convergencia cuadrática del método de Newton.

No obstante, al usar una fórmula de inversión matricial de Sherman y Morrison (véase, por ejemplo [DM, p. 55] se puede incorporar una mejora considerable.)

Teorema 10.8 **(Fórmula de Sherman-Morrison)**

Si A es una matriz no singular y si $\mathbf{x}$ y $\mathbf{y}$ son vectores, entonces $A + \mathbf{x}\mathbf{y}^t$ será no singular, a condición de que $\mathbf{y}^tA^{-1}\mathbf{x} \neq -1$, y

$$(A + \mathbf{x}\mathbf{y}^t)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{x}\mathbf{y}^t A^{-1}}{1 + \mathbf{y}^t A^{-1}\mathbf{x}}. \quad \blacksquare$$

Esta fórmula permite calcular A_i^{-1} directamente de A_{i-1}^{-1}, con lo cual se prescinde de una inversión matricial en cada iteración. Al utilizar $A = A_{i-1}$, $\mathbf{x} = (\mathbf{y}_i - A_{i-1}\mathbf{s}_i)/\|\mathbf{s}_i\|_2^2$ y $\mathbf{y} = \mathbf{s}_i$, la ecuación (10.13) junto con el teorema 10.8 implican que

$$A_i^{-1} = \left(A_{i-1} + \frac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{\|\mathbf{s}_i\|_2^2}\mathbf{s}_i^t\right)^{-1}$$

$$= A_{i-1}^{-1} - \frac{A_{i-1}^{-1}\left(\dfrac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{\|\mathbf{s}_i\|_2^2}\mathbf{s}_i^t\right)A_{i-1}^{-1}}{1 + \mathbf{s}_i^t A_{i-1}^{-1}\left(\dfrac{\mathbf{y}_i - A_{i-1}\mathbf{s}_i}{\|\mathbf{s}_i\|_2^2}\right)}$$

$$= A_{i-1}^{-1} - \frac{(A_{i-1}^{-1}\mathbf{y}_i - \mathbf{s}_i)\mathbf{s}_i^t A_{i-1}^{-1}}{\|\mathbf{s}_i\|_2^2 + \mathbf{s}_i^t A_{i-1}^{-1}\mathbf{y}_i - \|\mathbf{s}_i\|_2^2},$$

por lo que

$$A_i^{-1} = A_{i-1}^{-1} + \frac{(\mathbf{s}_i - A_{i-1}^{-1}\mathbf{y}_i)\mathbf{s}_i^t A_{i-1}^{-1}}{\mathbf{s}_i^t A_{i-1}^{-1}\mathbf{y}_i}. \qquad (10.16)$$

En este cálculo interviene exclusivamente la multiplicación de matrices y vectores en cada paso; por tanto, sólo se requieren $O(n^2)$ cálculos aritméticos. El cálculo de A_i se omite, y se prescinde de la resolución del sistema lineal (10.15). El algoritmo 10.2 se deduce directamente de esta construcción y se incorpora (10.16) al método iterativo (10.14).

ALGORITMO 10.2

Método de Broyden

Para aproximar la solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de ecuaciones e incógnitas; aproximación inicial $\mathbf{x} = (x_1, \ldots, x_n)^t$; tolerancia TOL; número máximo de iteraciones N.

SALIDA solución aproximada $\mathbf{x} = (x_1, \ldots, x_n)^t$ o mensaje de que se excedió el número de iteraciones.

Paso 1 Tome $A_0 = J(\mathbf{x})$, donde $J(\mathbf{x})_{i,j} = \frac{\partial f_i}{\partial x_j}(\mathbf{x})$ para $1 \leq i, j \leq n$;
$\mathbf{v} = \mathbf{F}(\mathbf{x})$. (*Nota*: $\mathbf{v} = \mathbf{F}(\mathbf{x}^{(0)})$.)

Paso 2 Tome $A = A_0^{-1}$. (*Utilice la eliminación gaussiana.*)

Paso 3 Tome $\mathbf{s} = -A\mathbf{v}$; (*Nota*: $\mathbf{s} = \mathbf{s}_1$.)
$\mathbf{x} = \mathbf{x} + \mathbf{s}$; (*Nota*: $\mathbf{x} = \mathbf{x}^{(1)}$.)
$k = 2$.

Paso 4 Mientras ($k \leq N$) haga los pasos 5–13.

Paso 5 Tome $\mathbf{w} = \mathbf{v}$; (*Guarde* $\mathbf{v}$.)
$\mathbf{v} = \mathbf{F}(\mathbf{x})$; (*Nota*: $\mathbf{v} = \mathbf{F}(\mathbf{x}^{(k)})$.)
$\mathbf{y} = \mathbf{v} - \mathbf{w}$. (*Nota*: $\mathbf{y} = \mathbf{y}_k$.)

Paso 6 Tome $\mathbf{z} = -A\mathbf{y}$. (*Nota*: $\mathbf{z} = A_{k-1}^{-1}\mathbf{y}_k$.)

Paso 7 Tome $p = -\mathbf{s}^t\mathbf{z}$. (*Nota*: $p = \mathbf{s}_k^t A_{k-1}^{-1}\mathbf{y}_k$.)

Paso 8 Tome $\mathbf{u}^t = \mathbf{s}^t A$.

Paso 9 Tome $A = A + \frac{1}{p}(\mathbf{s} + \mathbf{z})\mathbf{u}^t$. (*Nota*: $A = A_k^{-1}$.)

Paso 10 Tome $\mathbf{s} = -A\mathbf{v}$. (*Nota*: $\mathbf{s} = -A_k^{-1}\mathbf{F}(\mathbf{x}^{(k)})$.)

Paso 11 Tome $\mathbf{x} = \mathbf{x} + \mathbf{s}$. (*Nota*: $\mathbf{x} = \mathbf{x}^{(k+1)}$.)

Paso 12 Si $\|\mathbf{s}\| < TOL$, entonces SALIDA ($\mathbf{x}$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 13 Tome $k = k + 1$.

Paso 14 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

EJEMPLO 1 En el ejemplo 1 de la sección 10.2 resolvimos el sistema no lineal

$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0$$

aplicando el método de Newton. La matriz jacobiana para este sistema es

$$J(x_1, x_2, x_3) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_2 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}.$$

Con $\mathbf{x}^{(0)} = (0.1, 0.1, -0.1)^t$ tenemos

$$\mathbf{F}(x_1, x_2, x_3) = (f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3), f_3(x_1, x_2, x_3))^t,$$

donde

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - \frac{1}{2},$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06,$$

y

$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3}.$$

Entonces,

$$\mathbf{F}(\mathbf{x}^{(0)}) = \begin{bmatrix} -1.199950 \\ -2.269833 \\ 8.462025 \end{bmatrix}.$$

Dado que

$$A_0 = J(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})$$
$$= \begin{bmatrix} 3 & 9.999833 \times 10^{-4} & -9.999833 \times 10^{-4} \\ 0.2 & -32.4 & 0.9950042 \\ -9.900498 \times 10^{-2} & -9.900498 \times 10^{-2} & 20 \end{bmatrix},$$

tenemos

$$A_0^{-1} = J(x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^{-1}$$
$$= \begin{bmatrix} 0.3333332 & 1.023852 \times 10^{-5} & 1.615701 \times 10^{-5} \\ 2.108607 \times 10^{-3} & -3.086883 \times 10^{-2} & 1.535836 \times 10^{-3} \\ 1.660520 \times 10^{-3} & -1.527577 \times 10^{-4} & 5.000768 \times 10^{-2} \end{bmatrix}.$$

Por tanto,

$$\mathbf{x}^{(1)}, = \mathbf{x}^{(0)} - A_0^{-1}\mathbf{F}(\mathbf{x}^{(0)}) = \begin{bmatrix} 0.4998697 \\ 1.946685 \times 10^{-2} \\ -0.5215205 \end{bmatrix},$$

$$\mathbf{F}(\mathbf{x}^{(1)}) = \begin{bmatrix} -3.394465 \times 10^{-4} \\ -0.3443879 \\ 3.188238 \times 10^{-2} \end{bmatrix},$$

$$\mathbf{y}_1 = \mathbf{F}(\mathbf{x}^{(1)}) - \mathbf{F}(\mathbf{x}^{(0)}) = \begin{bmatrix} 1.199611 \\ 1.925445 \\ -8.430143 \end{bmatrix},$$

$$\mathbf{s}_1 = \begin{bmatrix} 0.3998697 \\ -8.053315 \times 10^{-2} \\ -0.4215204 \end{bmatrix},$$

$$\mathbf{s}^t_1 A_0^{-1}\mathbf{y}_i = 0.3424604,$$
$$A_1^{-1} = A_0^{-1} + (1/0.3424604)[(\mathbf{s}_1 - A_1^{-1}\,\mathbf{y}_1)\mathbf{s}^t_1\,A_0^{-1}]$$
$$= \begin{bmatrix} 0.3333781 & 1.11050 \times 10^{-5} & 8.967344 \times 10^{-6} \\ -2.021270 \times 10^{-3} & -3.094849 \times 10^{-2} & 2.196906 \times 10^{-3} \\ 1.022214 \times 10^{-3} & -1.650709 \times 10^{-4} & 5.010986 \times 10^{-2} \end{bmatrix},$$

y

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - A_1^{-1}\mathbf{F}(\mathbf{x}^{(1)}) = \begin{bmatrix} 0.4999863 \\ 8.737833 \times 10^{-3} \\ -0.5231746 \end{bmatrix}.$$

En la tabla 10.4 se incluyen otras iteraciones. La quinta iteración de Broyden es ligeramente menos precisa que la cuarta iteración de Newton, en el ejemplo dado al final de la sección anterior. ■

Tabla 10.4

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_2$
3	0.5000066	8.672157×10^{-4}	−0.5236918	7.88×10^{-3}
4	0.5000003	6.083352×10^{-2}	−0.5235954	8.12×10^{-4}
5	0.5000000	-1.448889×10^{-6}	−0.5235989	6.24×10^{-5}
6	0.5000000	6.059030×10^{-9}	−0.5235988	1.50×10^{-6}

También existen procedimientos para conservar la convergencia cuadrática, pero aminoran significativamente el número de evaluaciones funcionales que se requieren. Brown [Brow,K] fue el primero en proponer este tipo de métodos. En [MC] se encuentra una reseña y comparación de los métodos de este tipo de uso más común. Pero, en términos generales, son mucho más difíciles de implantar eficientemente que el método de Broyden.

CONJUNTO DE EJERCICIOS 10.3

1. Mediante el método de Broyden con $\mathbf{x}^{(0)} = \mathbf{0}$ calcule $\mathbf{x}^{(2)}$ para los siguientes sistemas no lineales.

   a. $4x_1^2 - 20x_1 + \frac{1}{4}x_2^2 + 8 = 0,$
   $\frac{1}{2}x_1x_2^2 + 2x_1 - 5x_2 + 8 = 0.$

   b. $\operatorname{sen}(4\pi x_1x_2) - 2x_2 - x_1 = 0,$
   $\left(\frac{4\pi - 1}{4\pi}\right)(e^{2x_1} - e) + 4ex_2^2 - 2ex_1 = 0.$

   c. $3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$
   $4x_1^2 - 625x_2^2 + 2x_2 - 1 = 0,$
   $e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$

   d. $x_1^2 + x_2 - 37 = 0,$
   $x_1 - x_2^2 - 5 = 0,$
   $x_1 + x_2 + x_3 - 3 = 0.$

2. Aplique el método de Broyden para aproximar las soluciones de los sistemas no lineales del ejercicio 1. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-6}$. Las aproximaciones iniciales $\mathbf{x}^{(0)}$ del ejercicio 1 quizá no lleven a la convergencia. De ser así, utilice otro valor de $\mathbf{x}^{(0)}$.

3. Con el método de Broyden, encuentre una solución a los siguientes sistemas no lineales. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-6}$.

   a. $3x_1^2 - x_2^2 = 0,$
   $3x_1x_2^2 - x_1^3 - 1 = 0.$
   Use $\mathbf{x}^{(0)} = (1, 1)^t$.

   b. $\ln(x_1^2 + x_2^2) - \operatorname{sen}(x_1x_2) = \ln 2 + \ln \pi,$
   $e^{x_1x_2} + \cos(x_1x_2) = 0.$
   Use $\mathbf{x}^{(0)} = (2, 2)^t$.

c. $x_1^3 + x_1^2 x_2 - x_1 x_3 + 6 = 0,$
$e^{x_1} + e^{x_2} - x_3 = 0,$
$x_2^2 - 2x_1 x_3 = 4.$
Use $\mathbf{x}^{(0)} = (-1, -2, 1)^t$.

d. $6x_1 - 2\cos(x_2 x_3) - 1 = 0,$
$9x_2 + \sqrt{x_1^2 + \operatorname{sen} x_3 + 1.06} + 0.9 = 0,$
$60x_3 + 3e^{-x_1 x_2} + 10\pi - 3 = 0.$
Use $\mathbf{x}^{(0)} = (0, 0, 0)^t$.

4. Aplique el método de Broyden para aproximar las soluciones a los siguientes sistemas no lineales. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-6}$.

a. $x_1(1 - x_1) + 4x_2 = 12,$
$(x_1 - 2)^2 + (2x_2 - 3)^2 = 25.$

b. $5x_1^2 - x_2^2 = 0,$
$x_2 - 0.25(\operatorname{sen} x_1 + \cos x_2) = 0.$

c. $15x_1 + x_2^2 - 4x_3 = 13,$
$x_1^2 + 10x_2 - x_3 = 11,$
$x_2^3 - 25x_3 = -22.$

d. $10x_1 - 2x_2^2 + x_2 - 2x_3 - 5 = 0,$
$8x_2^2 + 4x_3^2 - 9 = 0,$
$8x_2 x_3 + 4 = 0.$

5. El sistema no lineal

$$3x_1 - \cos(x_2 x_3) - \frac{1}{2} = 0,$$

$$x_1^2 - 625x_2^2 - \frac{1}{4} = 0,$$

$$e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

tiene una matriz jacobiana singular en la solución. Aplique el método de Broyden con $\mathbf{x}^{(0)} = (1, 1, -1)^t$. Observe que la convergencia puede ser lenta o no ocurrir dentro de un número razonable de iteraciones.

6. El sistema no lineal

$$\begin{aligned} 4x_1 - x_2 + x_3 &= x_1 x_4, \\ -x_1 + 3x_2 - 2x_3 &= x_2 x_4, \\ x_1 - 2x_2 + 3x_3 &= x_3 x_4, \\ x_1^2 + x_2^2 + x_3^2 &= 1 \end{aligned}$$

tiene seis soluciones.

a. Muestre que si $(x_1, x_2, x_3, x_4)^t$ es una solución, entonces $(-x_1, -x_2, -x_3, -x_4)^t$ también es una solución.

b. Use el método de Broyden tres veces para aproximar todas las soluciones. Itere hasta que $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 10^{-5}$.

7. El ejercicio 13 de la sección 8.1 se refiere a la determinación de la relación de mínimos cuadrados exponenciales de la forma $R = bw^a$ para aproximar un conjunto de datos concernientes al peso y a la regla de respiración de la polilla *Modest sphinx*. En ese ejercicio el problema se convirtió en una relación logaritmo-logaritmo, e introdujimos en la parte (c) un término cuadrático con el propósito de mejorar la aproximación. En vez de convertir el problema, determine las constantes a y b que reducen al mínimo $\sum_{i=1}^n (R_i - bw_i^a)^2$ para los datos del ejercicio 13 de la sección 8.1. Calcule el error relacionado con esta aproximación y después compárelo con el error de las aproximaciones anteriores en este problema.

8. Demuestre que si $\mathbf{0} \neq \mathbf{y} \in \mathbb{R}^n$ y si $\mathbf{z} \in \mathbb{R}^n$, entonces $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$, donde $\mathbf{z}_1 = (\mathbf{y}^t\mathbf{z}/\|\mathbf{y}\|_2^2)\mathbf{y}$ es paralelo a $\mathbf{y}$ y $\mathbf{z}_2$ es ortogonal a $\mathbf{y}$.

9. Demuestre que si $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, entonces $\det(I + \mathbf{u}\mathbf{v}^t) = 1 + \mathbf{v}^t\mathbf{u}$.

10. a. Use el resultado del ejercicio 9 para demostrar que si existe A^{-1} y $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$, entonces $(A + \mathbf{x}\mathbf{y}^t)^{-1}$ existe si y sólo si $\mathbf{y}^t A^{-1}\mathbf{x} \neq -1$.
   b. Al multiplicar a la derecha por $A + \mathbf{x}\mathbf{y}^t$, demuestre que, cuando $\mathbf{y}^t A^{-1}\mathbf{x} \neq -1$, tenemos

$$(A + \mathbf{x}\mathbf{y}^t)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{x}\mathbf{y}^t A^{-1}}{1 + \mathbf{y}^t A^{-1}\mathbf{x}}.$$

10.4 Métodos del descenso más rápido

La ventaja del método de Newton y de los métodos cuasi-Newton en la resolución de sistemas de ecuaciones no lineales es su rapidez de convergencia, una vez que se conoce una aproximación suficientemente exacta. Una de sus debilidades consiste en que requieren una aproximación inicial precisa de la solución para garantizar la convergencia. El método del **descenso más rápido** que estudiaremos en esta sección converge sólo linealmente a la solución, pero casi siempre convergirá incluso con aproximaciones iniciales deficientes. En consecuencia, con él se logran aproximaciones iniciales suficientemente exactas para las técnicas que tienen como base el método de Newton, del mismo modo que el método de la bisección se utiliza en una sola ecuación.

El método del descenso más rápido determina un mínimo local para una función de varias variables de la forma $g : \mathbb{R}^n \to \mathbb{R}$. Aunque el método es de gran utilidad independientemente de su aplicación como primer método para resolver los sistemas no lineales, limitaremos nuestra explicación a ese caso. (En los ejercicios se incluyen algunas otras aplicaciones.)

La conexión entre la minimización de una función de $\mathbb{R}^n$ a $\mathbb{R}$ y la solución de un sistema de ecuaciones no lineales se debe al hecho de que un sistema lineal de la forma

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ \vdots \qquad & \quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned}$$

tiene una solución en $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ justo cuando la función g definida por

$$g(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} [f_i(x_1, x_2, \ldots, x_n)]^2$$

tiene el valor mínimo cero.

El método del descenso más rápido para encontrar un mínimo local de una función arbitraria g de $\mathbb{R}^n$ a $\mathbb{R}$ puede describirse intuitivamente así:

1. Evalúe g en una aproximación inicial $\mathbf{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)^t$.
2. Determine una dirección desde $\mathbf{x}^{(0)}$ que origine una disminución del valor de g.
3. Desplace una cantidad apropiada hacia esta dirección y llame al nuevo vector $\mathbf{x}^{(1)}$.
4. Repita los pasos 1 a 3 reemplazando $\mathbf{x}^{(0)}$ con $\mathbf{x}^{(1)}$.

Antes de describir cómo seleccionar la dirección correcta y la distancia apropiada para desplazarse en esa dirección, es preciso revisar algunos resultados del cálculo. El teorema del valor extremo establece que una función diferenciable de una sola variable puede tener un mínimo relativo sólo cuando la derivada sea cero. Para extender este resultado a las funciones de varias variables necesitamos la siguiente definición.

Definición 10.9 Si $g : \mathbb{R}^n \to \mathbb{R}$, el **gradiente** de g en $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$ se denota con $\nabla g(\mathbf{x})$ y se define por medio de

$$\nabla g(\mathbf{x}) = \left(\frac{\partial g}{\partial x_1}(\mathbf{x}), \frac{\partial g}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial g}{\partial x_n}(\mathbf{x}) \right)^t. \qquad \blacksquare$$

El gradiente de una función de varias variables es análogo a la derivada de una función de una sola variable, en el sentido de que una función de varias variables diferenciable puede tener un mínimo relativo en $\mathbf{x}$ sólo cuando el gradiente sea cero.

El gradiente tiene otra propiedad muy importante relacionada con la minimización de las funciones de varias variables. Supóngase que $\mathbf{v} = (v_1, v_2, \ldots, v_n)^t$ es un vector unitario de $\mathbb{R}^n$; es decir,

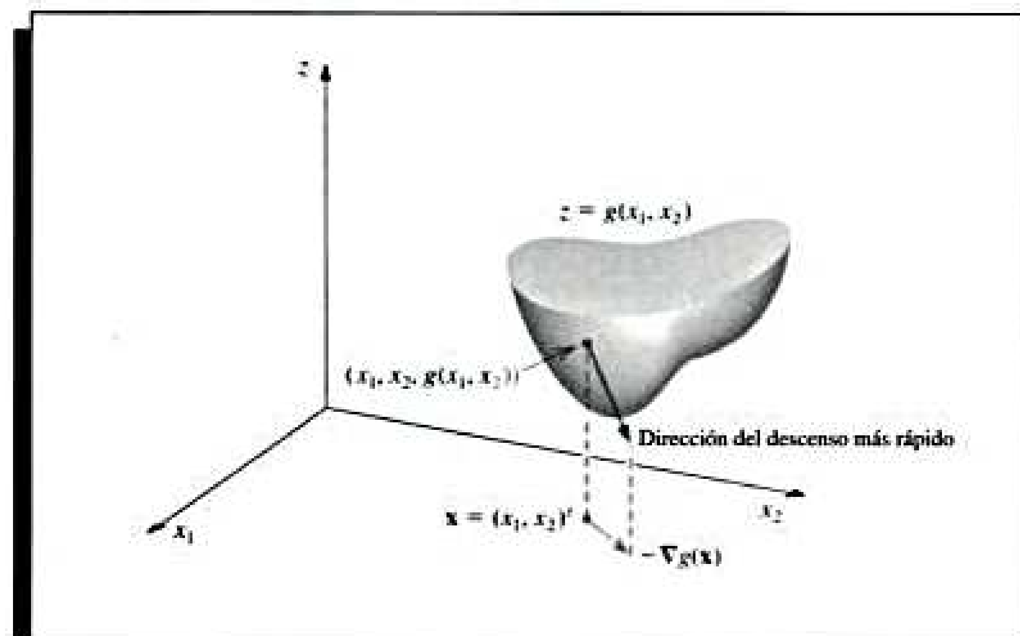
$$\|\mathbf{v}\|_2^2 = \sum_{i=1}^{n} v_i^2 = 1.$$

La **derivada direccional** de g en $\mathbf{x}$ en la dirección de $\mathbf{v}$ está definida por

$$D_{\mathbf{v}}g(\mathbf{x}) = \lim_{h \to 0} \frac{1}{h}[g(\mathbf{x} + h\mathbf{v}) - g(\mathbf{x})] = \mathbf{v}^t \cdot \nabla g(\mathbf{x}).$$

La derivada direccional de g en $\mathbf{x}$ en la dirección de $\mathbf{v}$ mide el cambio del valor de la función g respecto al cambio de la variable en la dirección de $\mathbf{v}$.

Figura 10.3



Un resultado estándar del cálculo de las funciones de varias variables establece que, si g es diferenciable, la dirección que produce el máximo valor para la derivada direccional se presenta cuando decidimos que $\mathbf{v}$ sea paralela a $\nabla g(\mathbf{x})$, siempre y cuando $\nabla g(\mathbf{x}) \neq \mathbf{0}$. En consecuencia, la dirección de la máxima disminución del valor de g en $\mathbf{x}$ es la dirección dada por $-\nabla g(\mathbf{x})$. (En la figura 10.3 se da un ejemplo de cuando g es una función de dos variables.)

Como la finalidad es reducir $g(\mathbf{x})$ a su valor mínimo cero, una elección apropiada de $\mathbf{x}^{(1)}$ será

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha \nabla g(\mathbf{x}^{(0)}), \quad \text{para alguna constante } \alpha > 0. \tag{10.17}$$

El problema se reduce así a escoger α tal que $g(\mathbf{x}^{(1)})$ sea significativamente menor que $g(\mathbf{x}^{(0)})$. Si queremos determinar una elección apropiada del valor α, consideramos la función de una sola variable

$$h(\alpha) = g(\mathbf{x}^{(0)} - \alpha \nabla g(\mathbf{x}^{(0)})). \tag{10.18}$$

El valor de α que minimiza h es el valor que se requiere en la ecuación (10.17).

Para obtener directamente un valor mínimo de h se requiere diferenciar h, y luego resolver un problema de cálculo de raíces para determinar los puntos críticos de h. Por lo general este procedimiento es costoso en términos de cálculos necesarios. Por ello seleccionamos tres números $\alpha_1 < \alpha_2 < \alpha_3$ que esperamos que estén cerca de donde ocurre el valor mínimo de $h(\alpha)$. Después construimos el polinomio cuadrático $P(x)$ que interpola h en α_1, α_2 y en α_3. Definimos $\hat{\alpha}$ en $[\alpha_1, \alpha_3]$ tal que $P(\hat{\alpha})$ sea un mínimo en $[\alpha_1, \alpha_3]$ y con $P(\hat{\alpha})$ aproximamos el valor mínimo de $h(\alpha)$. Luego usamos $\hat{\alpha}$ para determinar la nueva iteración con que aproximaremos el valor mínimo de g:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \hat{\alpha} \nabla g(\mathbf{x}^{(0)}).$$

Como $g(\mathbf{x}^{(0)})$ está disponible, primero escogemos $\alpha_1 = 0$ para disminuir en lo posible el cálculo necesario. A continuación encontramos un número α_3 con $h(\alpha_3) < h(\alpha_1)$. (Dado que α_1 no reduce al mínimo h, ese número α_3 sí existe. Finalmente decidimos que α_2 sea $\alpha_3/2$.

El valor mínimo de P en $[\alpha_1, \alpha_3]$ se presenta en el único punto crítico de P o en el punto extremo derecho α_3 porque, por suposición, $P(\alpha_3) = h(\alpha_3) < h(\alpha_1) = P(\alpha_1)$. Dado que P es un polinomio cuadrático, el punto crítico se puede determinar fácilmente.

EJEMPLO 1 Para encontrar una aproximación inicial razonable a la solución del sistema no lineal

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2 x_3) - \frac{1}{2} = 0,$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$f_3(x_1, x_2, x_3) = e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0,$$

utilizamos el método del descenso más rápido con $\mathbf{x}^{(0)} = (0, 0, 0)^t$.

Sea $g(x_1, x_2, x_3) = [f_1(x_1, x_2, x_3)]^2 + [f_2(x_1, x_2, x_3)]^2 + [f_3(x_1, x_2, x_3)]^2$; entonces

$$\nabla g(x_1, x_2, x_3) \equiv \nabla g(\mathbf{x}) = \left(2f_1(\mathbf{x}) \frac{\partial f_1}{\partial x_1}(\mathbf{x}) + 2f_2(\mathbf{x}) \frac{\partial f_2}{\partial x_1}(\mathbf{x}) + 2f_3(\mathbf{x}) \frac{\partial f_3}{\partial x_1}(\mathbf{x}),\right.$$

$$2f_1(\mathbf{x}) \frac{\partial f_1}{\partial x_2}(\mathbf{x}) + 2f_2(\mathbf{x}) \frac{\partial f_2}{\partial x_2}(\mathbf{x}) + 2f_3(\mathbf{x}) \frac{\partial f_3}{\partial x_2}(\mathbf{x}),$$

$$\left.2f_1(\mathbf{x}) \frac{\partial f_1}{\partial x_3}(\mathbf{x}) + 2f_2(\mathbf{x}) \frac{\partial f_2}{\partial x_3}(\mathbf{x}) + 2f_3(\mathbf{x}) \frac{\partial f_3}{\partial x_3}(\mathbf{x})\right)$$

$$= 2\mathbf{J}(\mathbf{x})^t\mathbf{F}(\mathbf{x}).$$

Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$, tenemos

$$g(\mathbf{x}^{(0)}) = 111.975 \quad \text{y} \quad z_0 = \|\nabla g(\mathbf{x}^{(0)})\|_2 = 419.554.$$

Sea

$$\mathbf{z} = \frac{1}{z_0} \nabla g(\mathbf{x}^{(0)}) = (-0.0214514, -0.0193062, 0.999583)^t.$$

Para $\alpha_1 = 0$, tenemos $g_1 = g(\mathbf{x}^{(0)} - \alpha_1\mathbf{z}) = g(\mathbf{x}^{(0)}) = 111.975$. De manera arbitraria, hacemos $\alpha_3 = 1$, de modo que

$$g_3 = g(\mathbf{x}^{(0)} - \alpha_3\mathbf{z}) = 93.5649.$$

Como $g_3 < g_1$, aceptamos α_3 y hacemos $\alpha_2 = 0.5$. Así,

$$g_2 = g(\mathbf{x}^{(0)} - \alpha_2\mathbf{z}) = 2.53557.$$

Ahora construimos el polinomio de interpolación de Newton con diferencias divididas hacia adelante

$$P(\alpha) = g_1 + h_1\alpha + h_3\alpha(\alpha - \alpha_2)$$

que interpola

$$g(\mathbf{x}^{(0)} - \alpha\nabla g(\mathbf{x}^{(0)})) = g(\mathbf{x}^{(0)} - \alpha\mathbf{z})$$

en $\alpha_1 = 0$, $\alpha_2 = 0.5$, y $\alpha_3 = 1$ como sigue:

$$\alpha_1 = 0, \quad g_1 = 111.975,$$

$$\alpha_2 = 0.5, \quad g_2 = 2.53557, \quad h_1 = \frac{g_2 - g_1}{\alpha_2 - \alpha_1} = -218.878,$$

$$\alpha_3 = 1 \quad g_3 = 93.5649, \quad h_2 = \frac{g_3 - g_2}{\alpha_3 - \alpha_2} = 182.059, \quad h_3 = \frac{h_2 - h_1}{\alpha_3 - \alpha_1} = 400.937.$$

Por tanto,

$$P(\alpha) = 111.975 - 218.878\alpha + 400.937\alpha(\alpha - 0.5).$$

Tenemos que $P'(\alpha) = 0$ cuando $\alpha = \alpha_0 = 0.522959$. Como $g_0 = g(\mathbf{x}^{(0)} - \alpha_0\mathbf{z}) = 2.32762$ es menor que g_1 y g_3, hacemos

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha_0\mathbf{z} = \mathbf{x}^{(0)} - 0.522959\mathbf{z} = (0.0112182, 0.0100964, -0.522741)^t$$

y

$$g(\mathbf{x}^{(1)}) = 2.32762.$$

La tabla 10.5 contiene el resto de los resultados. Una solución real del sistema no lineal es $(0.5, 0, -0.5235988)^t$. Aquí $\mathbf{x}^{(2)}$ será adecuado como aproximación inicial en los métodos de Newton y de Broyden. En esta etapa sería conveniente utilizar una de estas técnicas de convergencia más rápida, pues para obtener $\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty < 0.01$ se requieren 70 iteraciones del método del descenso más rápido. ■

Tabla 10.5

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$g(x_1^{(k)}, x_2^{(k)}, x_3^{(k)})$
2	0.137860	−0.205453	−0.522059	1.27406
3	0.266959	0.00551102	−0.558494	1.06813
4	0.272734	−0.00811751	−0.522006	0.468309
5	0.308689	−0.0204026	−0.533112	0.381087
6	0.314308	−0.0147046	−0.520923	0.318837
7	0.324267	−0.00852549	−0.528431	0.287024

En el algoritmo 10.3 se aplica el método del descenso más rápido para aproximar el valor mínimo de $g(\mathbf{x})$. Al comenzar una iteración, asignamos el valor 0 a α_1 y el valor 1 a α_3. Si $h(\alpha_3) \geq h(\alpha_1)$, entonces efectuamos divisiones consecutivas de α_3 entre 2 y reasignamos el valor de α_3 hasta que $h(\alpha_3) < h(\alpha_1)$ y $\alpha_3 = 2^{-k}$ para algún valor de k.

Si queremos emplear el método para aproximar la solución del sistema

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned}$$

tan sólo reemplazamos la función g con $\sum_{i=1}^n f_i^2$.

ALGORITMO 10.3

Método del descenso más rápido

Para aproximar una solución $\mathbf{p}$ al problema de minimización

$$g(\mathbf{p}) = \min_{\mathbf{x} \in \mathbb{R}^n} g(\mathbf{x})$$

dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de variables; aproximación inicial $\mathbf{x} = (x_1, \ldots, x_n)^t$; tolerancia TOL; número máximo de iteraciones N.

SALIDA solución aproximada $\mathbf{x} = (x_1, \ldots, x_n)^t$ o un mensaje de falla.

Paso 1 Tome $k = 1$.

Paso 2 Mientras ($k \leq N$), haga los pasos 3-15.

Paso 3 Tome $g_1 = g(x_1, \ldots, x_n)$; (*Nota*: $g_1 = g(\mathbf{x}^{(k)})$.)
$\mathbf{z} = \nabla g(x_1, \ldots, x_n)$; (*Nota*: $\mathbf{z} = \nabla g(\mathbf{x}^{(k)})$.)
$z_0 = \|\mathbf{z}\|_2$.

Paso 4 Si $z_0 = 0$, entonces SALIDA ('Gradiente cero');
SALIDA ($x_1, \ldots, x_n, g_1$);
(*Procedimiento terminado, puede tener un mínimo.*)
PARAR.

Paso 5 Tome $\mathbf{z} = \mathbf{z}/z_0$; (*Convierta a* $\mathbf{z}$ *en un vector unidad.*)
$\alpha_1 = 0$;
$\alpha_3 = 1$;
$g_3 = g(\mathbf{x} - \alpha_3\mathbf{z})$.

Paso 6 Mientras ($g_3 \geq g_1$), haga los pasos 7 y 8.

Paso 7 Tome $\alpha_3 = \alpha_3/2$;
$g_3 = g(\mathbf{x} - \alpha_3\mathbf{z})$.

Paso 8 Si $\alpha_3 < TOL/2$, entonces
SALIDA ('*Mejora poco probable*');
SALIDA ($x_1, \ldots, x_n, g_1$);
(*Procedimiento terminado, puede tener un mínimo.*)
PARAR.

Paso 9 Tome $\alpha_2 = \alpha_3/2$;
$g_2 = g(\mathbf{x} - \alpha_2\mathbf{z})$.

Paso 10 Tome $h_1 = (g_2 - g_1)/\alpha_2$;
$h_2 = (g_3 - g_2)/(\alpha_3 - \alpha_2)$;
$h_3 = (h_2 - h_1)/\alpha_3$.
(*Nota: La fórmula de las diferencias divididas hacia delante de Newton sirve para encontrar el cuadrático* $P(\alpha) = g_1 + h_1\alpha + h_3\alpha(\alpha - \alpha_2)$ *que interpola* $h(\alpha)$ *en* $\alpha = 0$, $\alpha = \alpha_2$, $\alpha = \alpha_3$.)

Paso 11 Tome $\alpha_0 = 0.5(\alpha_2 - h_1/h_3)$; (*El punto crítico de P ocurre en* α_0.)
$g_0 = g(\mathbf{x} - \alpha_0\mathbf{z})$.

Paso 12 Obtenga α de $\{\alpha_0, \alpha_3\}$ tal que $g = g(\mathbf{x} - \alpha\mathbf{z}) = \min\{g_0, g_3\}$.

Paso 13 Tome $\mathbf{x} = \mathbf{x} - \alpha\mathbf{z}$.

Paso 14 Si $|g - g_1| < TOL$, entonces
SALIDA ($x_1, \ldots, x_n, g$);
(*Procedimiento terminado exitosamente.*)
PARAR.

Paso 15 Tome $k = k + 1$.

Paso 16 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

El método del descenso más rápido admite muchas variaciones, algunas de las cuales incluyen técnicas más complejas para determinar el valor de α que producirá un mínimo con una función de una sola variable h definida en la ecuación (10.18). En otras técnicas se emplea el polinomio multidimensional de Taylor para reemplazar la función original g de varias variables y reducir al mínimo el polinomio en vez de g. Aunque algunas de ellas tienen ventajas sobre el procedimiento que hemos visto aquí, en general todos los métodos del descenso más rápido son linealmente convergentes y convergen independientemente de la aproximación inicial. Pero en algunos casos pueden convergir en algo que no es el mínimo absoluto de la función g.

En [OR] o en [RR] se explican con más detalle los métodos del descenso más rápido.

CONJUNTO DE EJERCICIOS 10.4

1. Aplique el método del descenso más rápido con $TOL = 0.05$ para aproximar las soluciones de los siguientes sistemas no lineales.

   a. $$4x_1^2 - 20x_1 + \frac{1}{4}x_2^2 + 8 = 0,$$
   $$\frac{1}{2}x_1x_2^2 + 2x_1 - 5x_2 + 8 = 0.$$

   b. $$3x_1^2 - x_2^2 = 0,$$
   $$3x_1x_2^2 - x_1^3 - 1 = 0.$$

   c. $$\ln(x_1^2 + x_2^2) - \operatorname{sen}(x_1x_2) = \ln 2 + \ln \pi,$$
   $$e^{x_1 - x_2} + \cos(x_1x_2) = 0.$$

   d. $$\operatorname{sen}(4\pi x_1x_2) - 2x_2 - x_1 = 0,$$
   $$\left(\frac{4\pi - 1}{4\pi}\right)(e^{2x_1} - e) + 4ex_2^2 - 2ex_1 = 0.$$

2. Use los resultados del ejercicio 1 y el método de Newton para aproximar, con una exactitud de 10^{-6}, los sistemas no lineales del ejercicio 1.

3. Aplique el método del descenso más rápido con $TOL = 0.05$ para aproximar las soluciones de los siguientes sistemas no lineales.

   a. $$15x_1 + x_2^2 - 4x_3 = 13,$$
   $$x_1^2 + 10x_2 - x_3 = 11,$$
   $$x_2^3 - 25x_3 = -22.$$

   b. $$10x_1 - 2x_2^2 + x_2 - 2x_3 - 5 = 0,$$
   $$8x_2^2 + 4x_3^2 - 9 = 0,$$
   $$8x_2x_3 + 4 = 0.$$

   c. $$x_1^3 + x_1^2x_2 - x_1x_3 + 6 = 0,$$
   $$e^{x_1} + e^{x_2} - x_3 = 0,$$
   $$x_2^2 - 2x_1x_3 = 4.$$

   d. $$x_1 + \cos(x_1x_2x_3) - 1 = 0,$$
   $$(1 - x_1)^{1/4} + x_2 + 0.05x_3^2 - 0.15x_3 - 1 = 0,$$
   $$-x_1^2 - 0.1x_2^2 + 0.01x_2 + x_3 - 1 = 0.$$

4. Use los resultados del ejercicio 3 y el método de Newton para aproximar, con una exactitud de 10^{-6}, las soluciones de los sistemas no lineales del ejercicio 3.

5. Aplique el método del descenso más rápido para aproximar, con una exactitud 0.005, los mínimos de las siguientes funciones.

   a. $g(x_1, x_2) = \cos(x_1 + x_2) + \operatorname{sen} x_1 + \cos x_2$

   b. $g(x_1, x_2) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2$

   c. $g(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + x_3^2 - 2x_1x_2 + 2x_1 - 2.5x_2 - x_3 + 2$

   a. $g(x_1, x_2, x_3) = x_1^4 + 2x_2^4 + 3x_3^4 + 1.01$

6. a. Demuestre que el polinomio cuadrático

$$P(\alpha) = g_1 + h_1\alpha + h_3\alpha(\alpha - \alpha_2)$$

interpola la función h definida en (10.18):

$$h(\alpha) = g(\mathbf{x}^{(0)} - \alpha\nabla g(\mathbf{x}^{(0)}))$$

en $\alpha = 0$, α_2 y α_3.

   b. Demuestre que un punto crítico de P ocurre en

$$\alpha_0 = \frac{1}{2}\left(\alpha_2 - \frac{h_1}{h_3}\right).$$

10.5 Métodos de homotopía y de continuación

Los métodos de *homotopía* o de *continuación* para sistemas no lineales introducen el problema por resolver dentro de una colección de problemas. Específicamente, para resolver un problema de la forma

$$\mathbf{F}(\mathbf{x}) = \mathbf{0},$$

con la solución desconocida $\mathbf{x}^*$, consideramos una familia de problemas descritos mediante un parámetro λ con valores en $[0, 1]$. Un problema con una solución conocida $\mathbf{x}^{(0)}$ corresponde a $\lambda = 0$, y el problema con la solución desconocida $\mathbf{x}^{(1)} \equiv \mathbf{x}^*$ corresponde a $\lambda = 1$.

Por ejemplo, supongamos que $\mathbf{x}(0)$ es una aproximación inicial de la solución de $\mathbf{F}(\mathbf{x}^*) = \mathbf{0}$. Definimos

$$\mathbf{G} : [0, 1] \times \mathbb{R}^n \to \mathbb{R}^n$$

como

$$\mathbf{G}(\lambda, \mathbf{x}) = \lambda\mathbf{F}(\mathbf{x}) + (1 - \lambda)[\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}(0))] = \mathbf{F}(\mathbf{x}) + (\lambda - 1)\mathbf{F}(\mathbf{x}(0)). \quad (10.19)$$

Para diversos valores de λ, determinaremos una solución de

$$\mathbf{G}(\lambda, \mathbf{x}) = \mathbf{0}.$$

Cuando $\lambda = 0$, esta ecuación asume la forma

$$\mathbf{0} = \mathbf{G}(0, \mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}(0)),$$

y $\mathbf{x}(0)$ es una solución. Cuando $\lambda = 1$, la ecuación asume la forma

$$\mathbf{0} = \mathbf{G}(1, \mathbf{x}) = \mathbf{F}(\mathbf{x}),$$

y $\mathbf{x}(1) = \mathbf{x}^*$ es una solución.

La función $\mathbf{G}$, con el parámetro λ, nos proporciona una familia de funciones que pueden conducir del valor conocido $\mathbf{x}(0)$ a la solución $\mathbf{x}(1) = \mathbf{x}^*$. La función $\mathbf{G}$ se llama una **homotopía** entre la función $\mathbf{G}(0, \mathbf{x}) = \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}(0))$ y la función $\mathbf{G}(1, \mathbf{x}) = \mathbf{F}(\mathbf{x})$.

El problema de **continuación** consiste en:

Determinar una forma de pasar de la solución conocida $\mathbf{x}(0)$ de $\mathbf{G}(0, \mathbf{x})$ a la solución desconocida $\mathbf{x}(1) = \mathbf{x}^*$ de $\mathbf{G}(1, \mathbf{x}) = \mathbf{0}$ que resuelva $\mathbf{F}(\mathbf{x}) = 0$.

Primero suponemos que $\mathbf{x}(\lambda)$ es la única solución de la ecuación

$$\mathbf{G}(\lambda, \mathbf{x}) = \mathbf{0}, \qquad (10.20)$$

para cada $\lambda \in [0, 1]$. El conjunto $\{\mathbf{x}(\lambda) \mid 0 \leq \lambda \leq 1\}$ se puede ver como una curva en $\mathbb{R}^n$ de $\mathbf{x}(0)$ a $\mathbf{x}(1) = \mathbf{x}^*$ parametrizada por λ. Un método de continuación determina una serie de pasos a lo largo de esta curva correspondiente a $\{\mathbf{x}(\lambda_k)\}_{k=0}^{m}$, donde $\lambda_0 = 0 < \lambda_1 < \cdots \lambda_m = 1$.

Si las funciones $\lambda \to \mathbf{x}(\lambda)$ y $\mathbf{G}$ son diferenciables, entonces al derivar la ecuación (10.20) con respecto de λ obtenemos

$$\mathbf{0} = \frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \lambda} + \frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \mathbf{x}} \mathbf{x}'(\lambda),$$

y al despejar $\mathbf{x}'(\lambda)$ se tiene

$$\mathbf{x}'(\lambda) = -\left[\frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \mathbf{x}}\right]^{-1} \frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \lambda}.$$

Este es un sistema de ecuaciones diferenciales con la condición inicial $\mathbf{x}(0)$.

Como

$$\mathbf{G}(\lambda, \mathbf{x}(\lambda)) = \mathbf{F}(\mathbf{x}(\lambda)) + (\lambda - 1)\mathbf{F}(\mathbf{x}(0)),$$

podemos determinar la matriz jacobiana

$$\frac{\partial \mathbf{G}}{\partial \mathbf{x}}(\lambda, \mathbf{x}(\lambda)) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}(\lambda)) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}(\lambda)) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}(\lambda)) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}(\lambda)) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}(\lambda)) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}(\lambda)) \\ \vdots & & & \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}(\lambda)) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}(\lambda)) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}(\lambda)) \end{bmatrix} = J(\mathbf{x}(\lambda)),$$

y

$$\frac{\partial \mathbf{G}(\lambda, \mathbf{x}(\lambda))}{\partial \lambda} = \mathbf{F}(\mathbf{x}(0)).$$

Por lo tanto, el sistema de ecuaciones diferenciales se convierte en

$$\mathbf{x}'(\lambda) = -[J(\mathbf{x}(\lambda))]^{-1}\mathbf{F}(\mathbf{x}(0)), \quad \text{para} \quad 0 \leq \lambda \leq 1, \qquad (10.21)$$

con la condición inicial $\mathbf{x}(0)$.

El siguiente teorema (véase [OR, pp. 230-231]) da condiciones bajo las cuales es factible el método de continuación.

Teorema 10.10 Sea $\mathbf{F}(\mathbf{x})$ una función continuamente diferenciable para $\mathbf{x} \in \mathbb{R}^n$. Suponga que la matriz jacobiana $J(\mathbf{x})$ es no singular para cada $\mathbf{x} \in \mathbb{R}^n$ y que existe una constante M tal que $\|J(\mathbf{x})^{-1}\| \leq M$, para cada $\mathbf{x} \in \mathbb{R}^n$. Entonces, para cualquier $\mathbf{x}(0)$ en $\mathbb{R}^n$, existe una única función $\mathbf{x}(\lambda)$, tal que

$$\mathbf{G}(\lambda, \mathbf{x}(\lambda)) = \mathbf{0},$$

para cada λ en $[0, 1]$. Además, $\mathbf{x}(\lambda)$ es continuamente diferenciable y

$$\mathbf{x}'(\lambda) = -J(\mathbf{x}(\lambda))^{-1}\mathbf{F}(\mathbf{x}(0)), \quad \text{para cada} \quad \lambda \in [0, 1]. \qquad \blacksquare$$

El siguiente ejemplo muestra la forma del sistema de ecuaciones diferenciales asociado con un sistema no lineal de ecuaciones.

EJEMPLO 1 Considere el sistema no lineal

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - 0.5 = 0,$$
$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$
$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0,$$

La matriz jacobiana es

$$J(\mathbf{x}) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_3 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}.$$

Sea $\mathbf{x}(0) = (0, 0, 0)^t$, de modo que

$$\mathbf{F}(\mathbf{x}(0)) = \begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}.$$

El sistema de ecuaciones diferenciales es

$$\begin{bmatrix} x_1'(\lambda) \\ x_2'(\lambda) \\ x_3'(\lambda) \end{bmatrix} = -\begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_2 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}^{-1} \begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}. \qquad \blacksquare$$

En general, el sistema de ecuaciones diferenciales que debemos resolver en nuestro problema de continuación tiene la forma

$$\frac{dx_1}{d\lambda} = \phi_1(\lambda, x_1, x_2, \ldots, x_n),$$
$$\frac{dx_2}{d\lambda} = \phi_2(\lambda, x_1, x_2, \ldots, x_n),$$
$$\vdots$$
$$\frac{dx_n}{d\lambda} = \phi_n(\lambda, x_1, x_2, \ldots, x_n),$$

donde

$$\begin{bmatrix} \phi_1(\lambda, x_1, \ldots, x_n), \\ \phi_2(\lambda, x_1, \ldots, x_n), \\ \vdots \\ \phi_n(\lambda, x_1, \ldots, x_n), \end{bmatrix} = -J(x_1, \ldots, x_n)^{-1} \begin{bmatrix} f_1(\mathbf{x}(0)) \\ f_2(\mathbf{x}(0)) \\ \vdots \\ f_n(\mathbf{x}(0)) \end{bmatrix}. \qquad (10.22)$$

Para resolver este sistema mediante el método de Runge-Kutta de cuarto orden, primero elegimos un entero $N > 0$ y hacemos $h = (1 - 0)/N$. Dividimos el intervalo $[0, 1]$ en N subintervalos con los puntos de red

$$\lambda_j = jh, \quad \text{para cada} \quad j = 0, 1, \ldots, N.$$

Usamos la notación w_{ij}, para cada $j = 0, 1, \ldots, N$ e $i = 1, \ldots, n$, para denotar una aproximación a $x_i(\lambda_j)$. Para la condición inicial, hacemos

$$w_{1,0} = x_1(0), \quad w_{2,0} = x_2(0), \quad w_{n,0} = x_n(0).$$

Suponga que hemos calculado $w_{1,j}, w_{2,j}, \ldots, w_{n,j}$. Obtenemos $w_{1,j+1}, w_{2,j+1}, \ldots, w_{n,j+1}$ usando las ecuaciones

$$k_{1,i} = h\phi_i(\lambda_j, w_{1,j}, w_{2,j}, \ldots, w_{n,j}), \quad \text{para cada} \ i = 1, 2, \ldots, n;$$

$$k_{2,i} = h\phi_i\left(\lambda_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{1,1}, w_{2,j} + \frac{1}{2}k_{1,1}, \ldots, w_{n,j} + \frac{1}{2}k_{1,n}\right),$$
$$\text{para cada} \quad i = 1, 2, \ldots, n;$$

$$k_{3,i} = h\phi_i\left(\lambda_j + \frac{h}{2}, w_{1,j} + \frac{1}{2}k_{2,1}, w_{2,j} + \frac{1}{2}k_{2,2}, \ldots, w_{n,j} + \frac{1}{2}k_{2,n}\right),$$
$$\text{para cada} \quad i = 1, 2, \ldots, n;$$

$$k_{4,i} = h\phi_i(\lambda_j + h, w_{1,j} + k_{3,1}, w_{2,j} + k_{3,2}, \ldots, w_{n,j} + k_{3,n}),$$
$$\text{para cada} \quad i = 1, 2, \ldots, n;$$

y, por último,

$$w_{i,j+1} = w_{1,j} + \frac{1}{6}(k_{1,i} + 2k_{2,i} + 2k_{3,i} + k_{4,i}), \quad \text{para cada} \ i = 1, 2, \ldots, n;$$

Usamos la notación vectorial

$$\mathbf{k}_1 = \begin{bmatrix} k_{1,1} \\ k_{1,2} \\ \vdots \\ k_{1,n} \end{bmatrix}, \quad \mathbf{k}_2 = \begin{bmatrix} k_{2,1} \\ k_{2,2} \\ \vdots \\ k_{2,n} \end{bmatrix}, \quad \mathbf{k}_3 = \begin{bmatrix} k_{3,1} \\ k_{3,2} \\ \vdots \\ k_{3,n} \end{bmatrix}, \quad \mathbf{k}_4 = \begin{bmatrix} k_{4,1} \\ k_{4,2} \\ \vdots \\ k_{4,n} \end{bmatrix} \quad \text{y} \quad \mathbf{w}_j = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{n,j} \end{bmatrix}$$

para simplificar la presentación.

La ecuación (10.22) implica que $\mathbf{x}(0) = \mathbf{x}(\lambda_0) = \mathbf{w}_0$, y para cada $j = 0, 1, \ldots, n$.

$$\mathbf{k}_1 = h\begin{bmatrix} \phi_1(\lambda_j, w_{1,j}, \ldots, w_{n,j}) \\ \phi_2(\lambda_j, w_{1,j}, \ldots, w_{n,j}) \\ \vdots \\ \phi_n(\lambda_j, w_{1,j}, \ldots, w_{n,j}) \end{bmatrix} = h\left[-J(\mathbf{w}_{1,j}, \ldots, w_{n,j})\right]^{-1}\mathbf{F}(\mathbf{x}(0))$$

$$= h\left[-J(\mathbf{w}_j)\right]^{-1}\mathbf{F}(\mathbf{x}(0));$$

$$\mathbf{k}_2 = h\left[-J\left(\mathbf{w}_j + \frac{1}{2}\mathbf{k}_1\right)\right]^{-1}\mathbf{F}(\mathbf{x}(0));$$

$$\mathbf{k}_3 = h\left[-J\left(\mathbf{w}_j + \frac{1}{2}\mathbf{k}_2\right)\right]^{-1}\mathbf{F}(\mathbf{x}(0));$$

$$\mathbf{k}_4 = h\left[-J(\mathbf{w}_j + \mathbf{k}_3)\right]^{-1}\mathbf{F}(\mathbf{x}(0));$$

y

$$\mathbf{x}(\lambda_{j+1}) = \mathbf{x}(\lambda_j) + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4) = \mathbf{w}_j + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4).$$

Por último, $\mathbf{x}(\lambda_n) = \mathbf{x}(1)$ es nuestra aproximación $\mathbf{x}^*$.

EJEMPLO 2 Aproximaremos la solución de

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - 0.5 = 0,$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 81(x_2 + 0.1)^2 + \operatorname{sen} x_3 + 1.06 = 0,$$

$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

La matriz jacobiana es

$$J(\mathbf{x}) = \begin{bmatrix} 3 & x_3 \operatorname{sen} x_2x_3 & x_2 \operatorname{sen} x_2x_3 \\ 2x_1 & -162(x_2 + 0.1) & \cos x_3 \\ -x_2e^{-x_1x_2} & -x_1e^{-x_1x_2} & 20 \end{bmatrix}$$

Sea $\mathbf{x}(0) = (0, 0, 0)^t$, de modo que

$$F(\mathbf{x}(0)) = (-1.5, 0.25, 10\pi/3)^t.$$

Con $N = 4$ y $h = 0.25$, tenemos

$$\mathbf{k}_1 = h[-J(\mathbf{x}^{(0)})]^{-1}F(\mathbf{x}(0)) = 0.25\begin{bmatrix} 3 & 0 & 0 \\ 0 & -16.2 & 1 \\ 0 & 0 & 20 \end{bmatrix}^{-1}\begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}$$

$$= (0.125, -0.004222203325, -0.1308996939)^t;$$

$$\mathbf{k}_2 = h[-J(0.0625, -0.002111101663, -0.06544984695)]^{-1}(-1.5, 0.25, 10\pi/3)^t$$

$$=0.25\begin{bmatrix} 3 & -0.9043289149\times 10^{-5} & -0.2916936196\times 10^{-6} \\ 0.125 & -15.85800153 & 0.9978589232 \\ 0.002111380229 & -0.06250824706 & 20 \end{bmatrix}^{-1}\begin{bmatrix} -1.5 \\ 0.25 \\ 10\pi/3 \end{bmatrix}$$

$$= (0.1249999773, -0.003311761993, -0.1309232406)^t;$$

$$\mathbf{k}_3 = h[-J(0.06249998865, -0.001655880997, -0.0654616203)]^{-1}(-1.5, 0.25, 10\pi/3)^t$$

$$= (0.1249999844, -0.003296244825, -0.130920346)^t;$$

$$\mathbf{k}_4 = h[-J(0.1249999844, -0.003296244825, -0.130920346)]^{-1}(-1.5, 0.25, 10\pi/3)^t$$

$$= (0.1249998945, -0.00230206762, -0.1309346977)^t;$$

y

$$\mathbf{x}(\lambda_1) = \mathbf{w}_1 = \mathbf{w}_0 + \frac{1}{6}[\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4]$$
$$= (0.1249999697, -0.00329004743, -0.1309202608)^t;$$

Continuando de esta forma, tenemos

$$\mathbf{x}(\lambda_2) = \mathbf{w}_2 = (0.2499997679, -0.004507400128, -0.2618557619)^t;$$
$$\mathbf{x}(\lambda_3) = \mathbf{w}_3 = (0.3749996956, -0.003430352103, -0.3927634423)^t;$$

y

$$\mathbf{x}(\lambda_4) = \mathbf{x}(1) = \mathbf{w}_4 = (0.4999999954, 0.126782\times 10^{-7}, -0.5235987758)^t.$$

Los resultados aquí obtenidos son muy precisos, puesta la solución real es aproximadamente $(0.5, 0, -0.52359877)^t$. ■

En el método de Runge-Kutta de orden cuatro, el cálculo de cada $\mathbf{w}_j$ requiere cuatro inversiones matriciales, una por cada vez que se calcula $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$ y $\mathbf{k}_4$. Así, el uso de N pasos requiere $4N$ inversiones matriciales. Por comparación, el método de Newton requiere una inversión matricial por iteración. Por tanto, el trabajo implicado en el método de Runge-Kutta es casi equivalente a $4N$ del método de Newton.

Una alternativa consiste en usar un método de Runge-Kutta de orden dos, como el método modificado de Euler o incluso el método de Euler, para disminuir el número de inversiones. Otra posibilidad de usar valores menores de N. El siguiente ejemplo ilustra estas ideas.

EJEMPLO 3 La tabla 10.6 resume una comparación de los métodos de Euler, del punto medio, y de Runge-Kutta de orden cuatro, aplicados al problema del ejemplo 2 con aproximación inicial $\mathbf{x}(0) = (0, 0, 0)^t$. ■

El método de continuación se puede usar como un método que no requiere una elección particularmente buena de $\mathbf{x}(0)$. Sin embargo, también se puede usar para obtener una aproximación inicial para los métodos de Newton o de Broyden. Por ejemplo, el resultado obtenido en el ejemplo 2 con el método de Euler y N = 2 podría ser suficiente

Tabla 10.6

Método	N	$\mathbf{x}(1)$	Número de inversiones
Euler	1	$(0.5, -0.0168888133, -0.5235987755)^t$	1
Euler	4	$(0.499999379, -0.004309160698, -0.523679652)^t$	4
Punto medio	1	$(0.4999966628, -0.00040240435, -0.523815371)^t$	2
Punto medio	4	$(0.500000066, -0.00001760089, -0.5236127761)^t$	8
Runge-Kutta	1	$(0.4999989843, -0.1676151 \times 10^{-5}, -0.5235989561)^t$	4
Runge-Kutta	4	$(0.4999999954, 0.126782 \times 10^{-7}, -0.5235987758)^t$	16

para iniciar cualquiera de los métodos más eficientes (Newton o Broyden) y para esto sería mejor que los métodos de continuación, que requieren más cálculos.

Observemos que en los métodos de Runge-Kutta, los pasos similares a

$$\mathbf{k}_i = h\,[-J(\mathbf{x}(\lambda_i) + \alpha_{i-1}\,\mathbf{k}_{i-1})]^{-1}\mathbf{F}(\mathbf{x}(0))$$

se pueden escribir como si se despejara $\mathbf{k}_i$ en el sistema lineal

$$\mathbf{J}\,(\mathbf{x}(\lambda_i) + \alpha_{i-1}\,\mathbf{k}_{i-1})\mathbf{k}_i = -h\mathbf{F}(\mathbf{x}(0)).$$

ALGORITMO 10.4

Algoritmo de continuación

Para aproximar la solución del sistema no lineal $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ dada una aproximación inicial $\mathbf{x}$:

ENTRADA número n de ecuaciones y de incógnitas; entero $N > 0$; aproximación inicial $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$.

SALIDA solución aproximada $\mathbf{x} = (x_1, x_2, \ldots, x_n)^t$.

Paso 1 Tome $h = 1/N$;
$\mathbf{b} = -h\mathbf{F}(\mathbf{x})$.

Paso 2 Para $i = 1, 2, \ldots, N$ haga los pasos 3-7.

Paso 3 Tome $A = J(\mathbf{x})$;
Resuelva el sistema lineal $A\mathbf{k}_1 = \mathbf{b}$.

Paso 4 Tome $A = J(\mathbf{x} + \frac{1}{2}\mathbf{k}_1)$;
Resuelva el sistema lineal $A\mathbf{k}_2 = \mathbf{b}$.

Paso 5 Tome $A = J(\mathbf{x} + \frac{1}{2}\mathbf{k}_2)$;
Resuelva el sistema lineal $A\mathbf{k}_3 = \mathbf{b}$.

Paso 6 Tome $A = J(\mathbf{x} + \mathbf{k}_3)$;
Resuelva el sistema lineal $A\mathbf{k}_4 = \mathbf{b}$.

Paso 7 Tome $\mathbf{x} = \mathbf{x} + (\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)/6$.

Paso 8 SALIDA $(x_1, x_2, \ldots, x_n)$;
PARAR. ■

CONJUNTO DE EJERCICIOS 10.5

1. El sistema no lineal

$$f_1(x_1, x_2) = x_1^2 - x_2^2 + 2x_2 = 0,$$
$$f_2(x_1, x_2) = 2x_1 + x_2^2 - 6 = 0$$

tiene dos soluciones $(0.625204094, 2.179355825)^t$ y $(2.109511920, -1.334532188)^t$. Aplique el método de continuación y el método de Euler con $N = 2$ para aproximar las soluciones, cuando

a. $\mathbf{x}(0) = (0, 0)^t$ **b.** $\mathbf{x}(0) = (1, 1)^t$ **c.** $\mathbf{x}(0) = (3, -2)^t$

2. Repita el ejercicio 1 usando el método de Runge-Kutta de orden cuatro con $N = 1$.
3. Aplique el método de continuación y el método de Euler con $N = 2$ a los siguientes sistemas no lineales.

a.
$$4x_1^2 - 20x_1 + \frac{1}{4}x_2^2 + 8 = 0,$$
$$\frac{1}{2}x_1x_2^2 + 2x_1 - 5x_2 + 8 = 0.$$

b.
$$\text{sen}\,(4\pi x_1x_2) - 2x_2 - x_1 = 0,$$
$$\left(\frac{4\pi - 1}{4\pi}\right)(e^{2x_1} - e) + 4ex_2^2 - 2ex_1 = 0.$$

c.
$$3x_1 - \cos(x_2x_3) - \frac{1}{2} = 0,$$
$$4x_1^2 - 625x_2^2 + 2x_2 - 1 = 0,$$
$$e^{-x_1x_2} + 20x_3 + \frac{10\pi - 3}{3} = 0.$$

d.
$$x_1^2 + x_2 - 37 = 0,$$
$$x_1 - x_2^2 - 5 = 0,$$
$$x_1 + x_2 + x_3 - 3 = 0.$$

4. Aplique el método de continuación y el método de Runge-Kutta de orden cuatro con $N = 1$ al ejercicio 4 de la sección 10.2 con $\mathbf{x}(0) = \mathbf{0}$. ¿Son comparables los resultados con los del citado ejercicio, o son aproximaciones iniciales adecuadas para el método de Newton?
5. Repita el ejercicio 4 usando la aproximación inicial obtenida en el ejercicio 2 de la sección 10.2.
6. Aplique el método de continuación y el método de Runge-Kutta de orden cuatro con $N = 1$ en el ejercicio 3 de la sección 10.2. ¿Son sus resultados tan buenos como los obtenidos entonces?
7. Repita el ejercicio 5 con $N = 2$.
8. Repita el ejercicio 6 de la sección 10.2 usando el método de continuación y el método de Runge-Kutta de orden cuatro con $N = 1$.
9. Repita el ejercicio 5 de la sección 10.2 usando el método de continuación y el método de Runge-Kutta de orden cuatro con $N = 2$.
10. Muestre que el método de continuación y el método de Euler con $N = 1$ dan el mismo resultado que el método de Newton para la primera iteración; es decir, si $\mathbf{x}(0) = \mathbf{x}^{(0)}$ siempre obtenemos $\mathbf{x}(1) = \mathbf{x}^{(1)}$.
11. Muestre que la homotopía

$$G(\lambda, \mathbf{x}) = F(\mathbf{x}) - e^{-\lambda} F(\mathbf{x}(0))$$

utilizada en el método de continuación con el método de Euler y $h = 1$ también duplica el método de Newton para cualquier $\mathbf{x}^{(0)}$; es decir, si $\mathbf{x}(0) = \mathbf{x}^{(0)}$, obtenemos $\mathbf{x}(1) = \mathbf{x}^{(1)}$.

12. Abreviaremos "método de continuación y el método de Runge-Kutta de orden cuatro" como CMRK4. Después de realizar los ejercicios 4, 5, 6, 7, 8 y 9, conteste lo siguiente.

a. ¿Es comparable CMRK4 con $N = 1$ con el método de Newton? Apoye su respuesta con los resultados de los ejercicios anteriores.

b. ¿Debemos usar CMRK4 $N = 1$ para obtener una aproximación inicial para el método de Newton? Apoye su respuesta con los resultados de los ejercicios anteriores.

c. Repita el inciso (a) para CMRK4 con $N = 2$.

d. Repita el inciso (b) para CMRK4 con $N = 2$.

10.6 Reseña de métodos y de software

En este capítulo estudiamos los métodos con que se aproximan las soluciones a los sistemas no lineales

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ f_2(x_1, x_2, \ldots, x_n) &= 0, \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0, \end{aligned}$$

El método de Newton para sistemas requiere una buena aproximación inicial $(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^t$ y genera la sucesión

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - J(\mathbf{x}^{(k-1)})^{-1}\,\mathbf{F}(\mathbf{x}^{(k-1)}),$$

que converge rápidamente en una solución $\mathbf{x}$ si $\mathbf{x}^{(0)}$ está suficientemente cerca de $\mathbf{p}$. Sin embargo, este método requiere evaluar o aproximar n^2 derivadas parciales y resolver un sistema lineal de $n \times n$ en cada paso, lo que requiere $O(n^3)$ cálculos.

El método de Broyden reduce la cantidad de cálculos en cada paso sin disminuir significativamente la rapidez de convergencia. Este método reemplaza a la matriz jacobiana J con una matriz A_{k-1} cuya inversa se determina directamente en cada paso. Así se reduce el número de cálculos aritméticos de $O(n^3)$ a $O(n^2)$. Además, las únicas evaluaciones funcionales escalares que se requieren son al evaluar f_i, lo cual constituye un ahorro de n^2 evaluaciones por paso. En el método de Broyden también se requiere una buena evaluación inicial.

Explicamos el método del descenso más rápido como una forma de obtener una buena aproximación inicial para los métodos de Newton y Broyden. Aunque el método del descenso más rápido no da una sucesión rápidamente convergente, no requiere una buena aproximación inicial; aproxima un mínimo de una función de varias variables g. En nuestra aplicación escogimos

$$g(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} [f_i(x_1, x_2, \ldots, x_n)]^2.$$

El mínimo de g es cero, que ocurre cuando las funciones f_i son simultáneamente cero.

La homotopía y los métodos de continuación también se emplean en los sistemas no lineales, y son tema de las investigaciones actuales. (Véase [AG].) En estos métodos, un problema dado

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}$$

está integrado en una familia de problemas de un parámetro que emplean un parámetro λ suponiendo valores en [0, 1]. El problema original corresponde a $\lambda = 1$, y un problema con una solución conocida corresponde a $\lambda = 0$. Por ejemplo, el conjunto de problemas

$$G(\lambda, \mathbf{x}) = \lambda\mathbf{F}(\mathbf{x}) + (1 - \lambda)(\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)) = \mathbf{0}, \quad \text{para } 0 \leq \lambda \leq 1,$$

para $\mathbf{x}_0 \in \mathbb{R}^n$ fijo, forma una homotopía. Cuando $\lambda = 0$, la solución es $\mathbf{x}(\lambda = 0) = \mathbf{x}_0$. La solución del problema original corresponde a $\mathbf{x}(\lambda = 1)$. Con un método de continuación se intenta determinar $\mathbf{x}(\lambda = 1)$ resolviendo la serie de problemas correspondientes a $\lambda_0 = 0 < \lambda_1 < \lambda_2 < ... < \lambda_m = 1$. La aproximación inicial de la solución de

$$\lambda_i\mathbf{F}(\mathbf{x}) + (1 - \lambda_i)(\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)) = \mathbf{0}$$

sería la solución, $\mathbf{x}(\lambda = \lambda_{i-1})$, del problema

$$\lambda_{i-1}\mathbf{F}(\mathbf{x}) + (1 - \lambda_{i-1})(\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_0)) = \mathbf{0}.$$

El paquete Hompack de Netlib resuelve un sistema de ecuaciones no lineales mediante varios métodos de homotopía.

En las bibliotecas IMSL y NAG los métodos de sistemas no lineales se basan en dos subrutinas IBRD y HYBRJ, contenidas en MINPACK, que es un paquete de dominio público. Ambos métodos utilizan la técnica de Levenberg-Marquardt, que es un promedio ponderado del método de Newton y del de descenso más rápido. La ponderación se inclina hacia el método del descenso rápido hasta que se descubre la convergencia; en ese momento el peso se inclina hacia el método de Newton, que converge más rápidamente. La subrutina HYBRD emplea una aproximación de diferencias finitas a la matriz jacobiana; para calcular la matriz jacobiana HYBRJ se requiere una subrutina proporcionada por el usuario.

Con la subrutina NEQNF de IMSL se resuelve un sistema no lineal, sin que el usuario tenga que introducir una matriz jacobiana. La subrutina NEQNJ se parece a la anterior, salvo que el usuario debe introducir una subrutina para calcular la matriz jacobiana.

En la biblioteca NAG, CO5NBF se parece a HYBRD. La subrutina CO5PBF se parece a CO5NBF, excepto que el usuario debe introducir una subrutina para calcular la matriz jacobiana. La subrutina CO5PBF se basa en HYBRJ del paquete MINPACK. NAG también contiene otras modificaciones del método de Levenberg-Marquardt.

Un tratamiento muy exhaustivo de los métodos con que se resuelven los sistemas no lineales de ecuaciones se encuentra en Ortega y Rheinbolt [OR], y en Dennis y Schnabel [DenS]. Los avances recientes en los métodos iterativos se explican en Argyros y Szidarovszky [AS]; el lector que desee información sobre el empleo de los métodos continuos puede consultar a Allgower y Georg [AG].

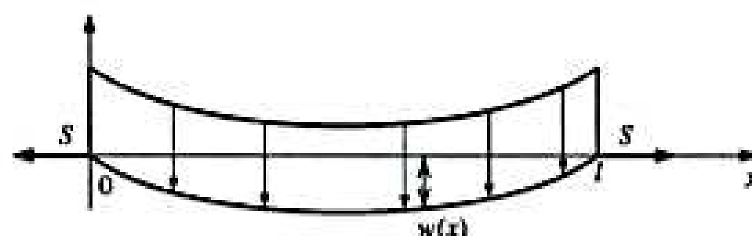
CAPÍTULO

11

Problemas con valor en la frontera para ecuaciones diferenciales ordinarias

. . .

Un problema común en ingeniería civil es el que se relaciona con la deflexión de una viga de sección transversal rectangular sujeta a una carga uniforme, mientras sus extremos están soportados de modo que no experimentan deflexión alguna.

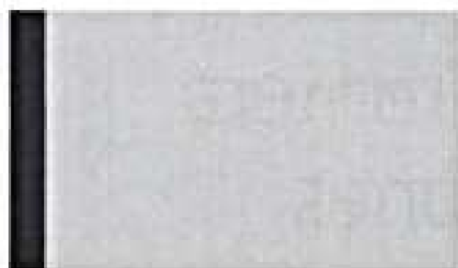


La ecuación diferencial que aproxima la situación física es de la forma

$$\frac{d^2w}{dx^2}(x) = \frac{S}{EI}w(x) + \frac{qx}{2EI}(x - l),$$

donde $w(x)$ es la deflexión a una distancia x desde el extremo izquierdo de la viga, y l, q, E, S e I representan, respectivamente, la longitud de la viga, la intensidad de la carga uniforme, el módulo de elasticidad, el esfuerzo en los extremos y el momento central de inercia. Esta ecuación diferencial tiene asociadas dos condiciones de frontera dadas por la suposición de que no ocurre deflexión alguna en los extremos de la viga

$$w(0) = w(l) = 0.$$



Cuando la viga tiene un espesor uniforme, el producto EI es constante y la solución exacta se obtiene fácilmente. No obstante, en muchas aplicaciones el espesor no es uniforme y, por tanto, el momento de inercia I es una función de x, y se requieren métodos de aproximación. Este tipo de problemas se consideran en el ejercicio 7 de la sección 11.3 y el 6 de la sección 11.4.

Las ecuaciones diferenciales del capítulo 5 eran de primer orden y debían satisfacer una condición inicial. Más adelante, en dicho capítulo vimos que las técnicas podían extenderse a sistemas de ecuaciones y luego a ecuaciones de orden superior, pero todas las condiciones fueron dadas en el mismo punto extremo. Éstos son problemas de valor inicial. En este capítulo mostraremos cómo aproximar la solución de problemas **con valores de frontera**, ecuaciones diferenciales con condiciones impuestas en distintos puntos. Para las ecuaciones diferenciales de primer orden, sólo se especifica una condición, así que no hay distinción entre los problemas con valor inicial y con valor de frontera. Aquí analizaremos ecuaciones de segundo orden con dos valores de frontera.

Los problemas físicos que dependen más de la posición que del tiempo a menudo se describen en función de ecuaciones diferenciales, con las condiciones impuestas en más de un punto. En este capítulo, los problemas con valor en la frontera de dos puntos incluyen una ecuación diferencial de segundo orden de la forma

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \tag{11.1}$$

junto con las condiciones de frontera

$$y(a) = \alpha \quad \text{y} \quad y(b) = \beta. \tag{11.2}$$

11.1 El método del disparo lineal

El siguiente teorema establece las condiciones generales que garantizan que exista la solución a un problema con valor en la frontera de segundo orden y que dicha solución sea única. La prueba del teorema se encuentra en [Keller, H].

Teorema 11.1 Supongamos que la función f en el problema con valor en la frontera

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

es continua en el conjunto

$$D = \{(x, y, y') \mid a \leq x \leq b, -\infty < y < \infty, -\infty < y' < \infty\},$$

y que f_y y $f_{y'}$ también son continuas en D. Si

(i) $f_y(x, y, y') > 0$ para toda $(x, y, y') \in D$, y

(ii) existe una constante M, con

$$|f_{y'}(x, y, y')| \leq M, \quad \text{para toda} \quad (x, y, y') \in D,$$

entonces el problema con valor en la frontera tiene una solución única. ■

EJEMPLO 1 El problema con valor en la frontera

$$y'' + e^{-xy} + \text{sen}\, y' = 0, \quad 1 \leq x \leq 2, \quad y(1) = y(2) = 0,$$

tiene

$$f(x, y, y') = -e^{-xy} - \text{sen}\, y'.$$

Puesto que

$$f_y(x, y, y') = xe^{-xy} > 0 \quad \text{y} \quad |f_{y'}(x, y, y')| = |-\cos y'| \leq 1,$$

este problema tiene una solución única. ■

Cuando $f(x, y, y')$ tiene la forma

$$f(x, y, y') = p(x)y' + q(x)y + r(x),$$

la ecuación diferencial

$$y'' = f(x, y, y')$$

es **lineal**. Este tipo de problemas ocurren frecuentemente, y en este caso el teorema 11.1 puede simplificarse.

Corolario 11.2 Si el problema lineal con valor en la frontera

$$y'' = p(x)y' + q(x)y + r(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

satisface

(i) $p(x)$, $q(x)$ y $r(x)$ son continuas en $[a, b]$,
(ii) $q(x) > 0$ en $[a, b]$,

entonces el problema tiene una solución única. ■

Para aproximar la solución única garantizada por el cumplimiento de las hipótesis del corolario 11.2, primero consideraremos los problemas con valor inicial

$$y'' = p(x)y' + q(x)y + r(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y'(a) = 0, \tag{11.3}$$

y

$$y'' = p(x)y' + q(x)y, \quad a \leq x \leq b, \quad y(a) = 0, \quad y'(a) = 1. \tag{11.4}$$

El teorema 5.16 de la sección 5.9 garantiza que, según las hipótesis del corolario 11.2, ambos problemas tienen una solución única. Si $y_1(x)$ denota la solución de (11.3) y si $y_2(x)$ denota la solución de (11.4), no es difícil comprobar que

$$y(x) = y_1(x) + \frac{\beta - y_1(b)}{y_2(b)}\, y_2(x). \tag{11.5}$$

Entonces

$$y'(x) = y_1'(x) + \frac{\beta - y_1(b)}{y_2(b)}\, y_2'(x)$$

y

$$y''(x) = y_1''(x) + \frac{\beta - y_1(b)}{y_2(b)}\, y_2''(x).$$

Por tanto,

$$\begin{aligned} y'' &= p(x)\, y_1' + q(x) y_1 + r(x) + \frac{\beta - y_1(b)}{y_2(b)} \left(p(x) y_2' + q(x) y_2\right) \\ &= p(x)\left(y_1' + \frac{\beta - y_1(b)}{y_2(b)} y_2'\right) + q(x)\left(y_1 + \frac{\beta - y_1(b)}{y_2(b)} y_2\right) + r(x) \\ &= p(x) y'(x) + q(x)\, y(x) + r(x). \end{aligned}$$

Más aún

$$y(a) = y_1(a) + \frac{\beta - y_1(b)}{y_2(b)}\, y_2(a) = \alpha + \frac{\beta - y_1(b)}{y_2(b)} \cdot 0 = \alpha$$

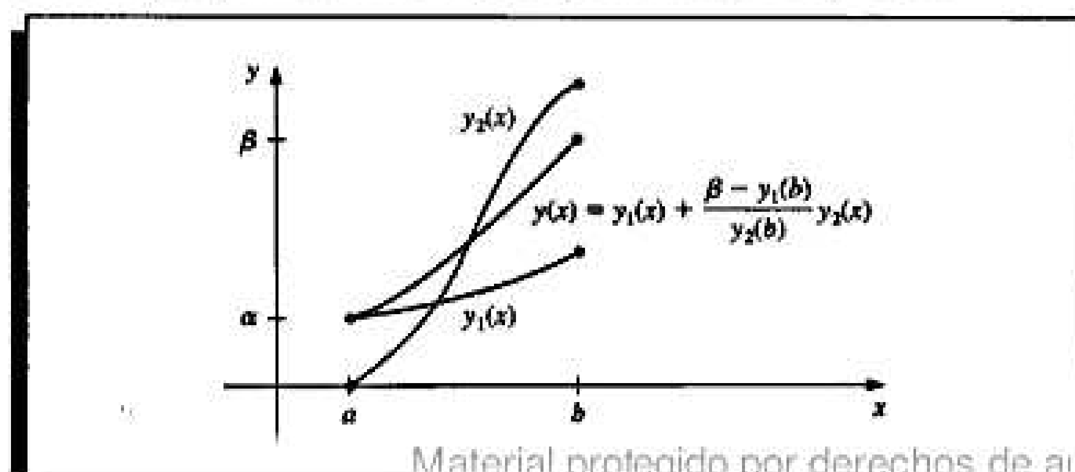
y

$$y(b) = y_1(b) + \frac{\beta - y_1(b)}{y_2(b)} y_2(b) = y_1(b) + \beta - y_1(b) = \beta.$$

Por tanto, $y(x)$ es la solución única a nuestro problema con valor en la frontera, naturalmente a condición de que $y_2(b) \neq 0$. (En el ejercicio 8 se considera que $y_2(b) = 0$ choca con las hipótesis del corolario 11.2.)

El método del disparo para las ecuaciones lineales se basa en la sustitución del problema lineal con valor en la frontera por dos problemas con valor inicial (11.3) y (11.4). En el capítulo 5 se describen muchos métodos con los cuales podemos aproximar las soluciones $y_1(x)$ y $y_2(x)$, y una vez que contamos con estas aproximaciones, la solución del problema de valor de frontera se aproxima por medio de la ecuación (11.5). Desde el punto de vista gráfico, el método tiene el aspecto que se observa en la figura 11.1.

Figura 11.1



En el algoritmo 11.1 se usa el método de Runge-Kutta de cuarto orden para obtener las aproximaciones a $y_1(x)$ y a $y_2(x)$, pero en el paso 4 puede sustituirse cualquier otra técnica con que se aproximen las soluciones a los problemas de valor inicial.

El algoritmo tiene la característica adicional de obtener aproximaciones para la derivada de la solución del problema con valor de frontera y también de la solución del problema en sí mismo. El uso del algoritmo no se limita a los problemas en que se puede verificar el corolario 11.2; también da resultados satisfactorios en muchos problemas que no satisfacen esas hipótesis.

ALGORITMO 11.1

Método del disparo lineal

Para aproximar la solución del problema con valor de frontera

$$-y''+ p(x)y' + q(x)y + r(x) = 0, \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta:$$

(*Nota: las ecuaciones* 11.3 *y* 11.4 *se escriben y se resuelven como sistemas de primer orden.*)

ENTRADA extremos a, b; condiciones de frontera α, β; número de subintervalos N.

SALIDA aproximaciones $w_{1,i}$ a $y(x_i)$; $w_{2,i}$ a $y'(x_i)$ para cada $i = 0, 1, \ldots, N$.

Paso 1 Tome $h = (b - a)/N$;
$u_{1,0} = \alpha$;
$u_{2,0} = 0$;
$v_{1,0} = 0$;
$v_{2,0} = 1$.

Paso 2 Para $i = 0, \ldots, N - 1$ haga los pasos 3-4.
(*El método Runge-Kutta para sistemas se utiliza en pasos* 3 *y* 4.)

Paso 3 Tome $x = a + ih$.

Paso 4 Tome $k_{1,1} = hu_{2,i}$;

$k_{1,2} = h\left[p(x)u_{2,i} + q(x)u_{1,i} + r(x)\right]$;

$k_{2,1} = h\left[u_{2,i} + \frac{1}{2}k_{1,2}\right]$;

$k_{2,2} = h\left[p(x + h/2)\left(u_{2,i} + \frac{1}{2}k_{1,2}\right) + q(x + h/2)\left(u_{1,i} + \frac{1}{2}k_{1,1}\right) + r(x + h/2)\right]$;

$k_{3,1} = h\left[u_{2,i} + \frac{1}{2}k_{2,2}\right]$;

$k_{3,2} = h\left[p(x + h/2)\left(u_{2,i} + \frac{1}{2}k_{2,2}\right) + q(x + h/2)\left(u_{1,i} + \frac{1}{2}k_{2,1}\right) + r(x + h/2)\right]$;

$k_{4,1} = h\left[u_{2,i} + k_{3,2}\right]$;

$k_{4,2} = h\left[p(x + h)\,(u_{2,i} + k_{3,2}) + q(x + h)(u_{1,i} + k_{3,1}) + r(x + h)\right]$;

$u_{1,i+1} = u_{1,i} + \frac{1}{6}\left[k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1}\right]$;

$u_{2,i+1} = u_{2,i} + \frac{1}{6}\left[k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2}\right]$;

$k'_{1,1} = hv_{2,i}$;

$k'_{1,2} = h\left[p(x)u_{2,i} + q(x)v_{1,i}\right];$
$k'_{2,1} = h\left[u_{2,i} + \frac{1}{2}k'_{1,2}\right];$
$k'_{2,2} = h\left[p(x + h/2)\left(u_{2,i} + \frac{1}{2}k'_{1,2}\right) + q(x + h/2)(v_{1,i} + \frac{1}{2}k'_{1,1})\right];$
$k'_{3,1} = h\left[u_{2,i} + \frac{1}{2}k'_{2,2}\right];$
$k'_{3,2} = h\left[p(x + h/2)\left(u_{2,i} + \frac{1}{2}k'_{2,2}\right) + q(x + h/2)\left(v_{1,i} + \frac{1}{2}k'_{2,1}\right)\right];$
$k'_{4,1} = h\left[v_{2,i} + k'_{3,2}\right];$
$k'_{4,2} = h\left[p(x + h)(u_{2,i} + k'_{3,2}) + q(x + h)(v_{1,i} + k'_{3,1})\right];$
$v_{1,i+1} = v_{1,i} + \frac{1}{6}\left[k'_{1,1} + 2k'_{2,1} + 2k'_{3,1} + k'_{4,1}\right];$
$v_{2,i+1} = v_{2,i} + \frac{1}{6}\left[k'_{1,2} + 2k'_{2,2} + 2k'_{3,2} + k'_{4,2}\right].$

Paso 5 Tome $w_{1,0} = \alpha;$

$$w_{2,0} = \frac{\beta - u_{1,N}}{v_{1,N}};$$

SALIDA $(a, w_{1,0}, w_{2,0})$.

Paso 6 Para $i = 1, \ldots, N$

tome $W1 = u_{1,i} + w_{2,0}\, v_{1,i};$
$W2 = u_{2,i} + w_{2,0}\, v_{2,i};$
$x = a + ih;$
SALIDA $(x, W1, W2)$. (*La salida es* $x_i,\ w_{1,i},\ w_{2,i}$)

Paso 7 PARAR. (*Procedimiento terminado.*) ■

EJEMPLO 2 El problema con valor de frontera

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad 1 \le x \le 2, \quad y(1) = 1, \quad y(2) = 2,$$

tiene la solución exacta

$$y = c_1x + \frac{c_2}{x^2} - \frac{3}{10}\operatorname{sen}(\ln x) - \frac{1}{10}\cos(\ln x),$$

donde

$$c^2 = \frac{1}{70}\left[8 - 12\operatorname{sen}(\ln 2) - 4\cos(\ln 2) \approx -0.03920701320\right.$$

y

$$c_1 = \frac{11}{10} - c_2 \approx 1.1392070132.$$

Si queremos aplicar el algoritmo 11.1 a este problema, es necesario aproximar las soluciones de los problemas con valor inicial

$$y_1'' = -\frac{2}{x}y_1' + \frac{2}{x^2}y_1 + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad 1 \le x \le 2, \quad y_1(1) = 1, \quad y_1'(1) = 0.$$

y

$$y_2'' = -\frac{2}{x}y_2' + \frac{2}{x^2}y_2, \quad 1 \le x \le 2, \quad y_2(1) = 0, \quad y_2'(1) = 1.$$

En la tabla 11.1 se incluyen los resultados de los cálculos cuando se emplea el algoritmo 11.1 con $N = 10$ y $h = 0.1$. El valor marcado como $u_{1,i}$ aproxima $y_1(x_i)$, $v_{1,i}$ aproxima $y_2(x_i)$ y w_i aproxima

$$y(x_i) = y_1(x_i) + \frac{2 - y_1(2)}{y_2(2)} y_2(x_i). \quad \blacksquare$$

Tabla 11.1

x_i	$u_{1,i}$	$v_{1,i}$	w_i	$y(x_i)$	$\lvert y(x_i) - w_i \rvert$
1.0	1.00000000	0.00000000	1.00000000	1.00000000	
1.1	1.00896058	0.09117986	1.09262917	1.09262930	1.43×10^{-7}
1.2	1.03245472	0.16851175	1.18708471	1.18708484	1.34×10^{-7}
1.3	1.06674375	0.23608704	1.28338227	1.28338236	9.78×10^{-8}
1.4	1.10928795	0.29659067	1.38144589	1.38144595	6.02×10^{-8}
1.5	1.15830000	0.35184379	1.48115939	1.48115942	3.06×10^{-8}
1.6	1.21248372	0.40311695	1.58239245	1.58239246	1.08×10^{-8}
1.7	1.27087454	0.45131840	1.68501396	1.68501396	5.43×10^{-10}
1.8	1.33273851	0.49711137	1.78889854	1.78889853	5.05×10^{-9}
1.9	1.39750618	0.54098928	1.89392951	1.89392951	4.41×10^{-9}
2.0	1.46472815	0.58332538	2.00000000	2.00000000	

La exactitud que muestra la tabla 11.1 es previsible porque el método de Runge-Kutta de cuarto orden ofrece una exactitud $O(h^4)$ a las soluciones de los problemas de valor inicial. Desafortunadamente, esta técnica, por errores de redondeo, puede contener problemas ocultos. Si $y_1(x)$ aumenta rápidamente a medida que x pasa de a a b, entonces $u_{1,N} \approx y_1(b)$ será grande. Si β tiene una magnitud pequeña en comparación con $u_{1,N}$, el término $w_{2,0} = (\beta - u_{1,N})/v_{1,N}$ será aproximadamente $-u_{1,N}/v_{1,N}$. Entonces los cálculos del paso 6 se convierten en

$$W1 = u_{1,i} + w_{2,0}v_{1,i} \approx u_{1,i} - \left(\frac{u_{1,N}}{v_{1,N}}\right) v_{1,i},$$

$$W2 = u_{2,i} + w_{2,0}v_{2,i} \approx u_{2,i} - \left(\frac{u_{1,N}}{v_{1,N}}\right) v_{2,i},$$

lo cual permite una posible pérdida de los dígitos significativos debido a la cancelación. Pero como $u_{1,i}$ es una aproximación a $y_1(x_i)$, podemos vigilar fácilmente el comportamiento de y_1, y si $u_{1,i}$ aumenta rápidamente de a a b, podemos aplicar hacia atrás el método del disparo, esto es, resolver en su lugar los problemas de valor inicial

$$y'' = p(x)y' + q(x)y + r(x), \quad a \le x \le b, \quad y(b) = \beta, \quad y'(b) = 0,$$

y

$$y'' = p(x)y' + q(x)y, \quad a \le x \le b, \quad y(b) = 0, \quad y'(b) = 1.$$

Si este método del disparo inverso todavía ofrece la eliminación de los dígitos significativos y si el aumento de precisión no produce mayor exactitud, será necesario utilizar otras técnicas como las que explicaremos más adelante en este capítulo. No obstante, en general si $u_{1,i}$ y $v_{1,i}$ son aproximaciones de $O(h^n)$ para $y_1(x_i)$ y $y_2(x_i)$ respectivamente, con $i = 0, 1, \ldots, N$, entonces $w_{1,i}$ será una aproximación de $O(h^n)$ para $y(x_i)$. En particular,

$$|w_{1,i} - y(x_i)| \leq Kh^n \left| 1 + \frac{v_{1,i}}{v_{1,N}} \right|,$$

para alguna constante K (véase [IK, p. 426]).

CONJUNTO DE EJERCICIOS 11.1

1. El problema con valor de frontera

$$y'' = 4(y - x), \quad 0 \leq x \leq 1, \quad y(0) = 0, \quad y(1) = 2,$$

tiene la solución $y(x) = e^2(e^4 - 1)^{-1}(e^{2x} - e^{-2x}) + x$. Aplique el método del disparo lineal para aproximar la solución, y después compare los resultados con la solución real.

a. Con $h = \frac{1}{2}$; b. Con $h = \frac{1}{4}$.

2. El problema con valor de frontera

$$y'' = y' + 2y + \cos x, \quad 0 \leq x \leq \frac{\pi}{2}, \quad y(0) = -0.3, \quad y\left(\frac{\pi}{2}\right) = -0.1$$

tiene la solución $y(x) = -\frac{1}{10}(\operatorname{sen} x + 3\cos x)$. Aplique el método del disparo lineal para aproximar la solución, y después compare los resultados con la solución real.

a. Con $h = \frac{\pi}{4}$; b. Con $h = \frac{\pi}{8}$.

3. Aplique el método del disparo lineal para aproximar la solución a los siguientes problemas con valor de frontera.

a. $y'' = -3y' + 2y + 2x + 3, \quad 0 \leq x \leq 1, \quad y(0) = 2, y(1) = 1$; use $h = 0.1$.
b. $y'' = -\frac{4}{x}y' + \frac{2}{x^2}y - \frac{2}{x^2}\ln x, \quad 1 \leq x \leq 2, \quad y(1) = -\frac{1}{2}, \quad y(2) = \ln 2$; use $h = 0.05$.
c. $y'' = -(x + 1)y' + 2y + (1 - x^2)e^{-x}, \quad 0 \leq x \leq 1, \quad y(0) = -1, y(1) = 0$; use $h = 0.1$.
d. $y'' = \frac{y'}{x} + \frac{3}{x^2}y + \frac{\ln x}{x} - 1, \quad 1 \leq x \leq 2, \quad y(1) = y(2) = 0$; use $h = 0.1$.

4. Aunque $q(x) < 0$ en los siguientes problemas con valor de frontera, existen soluciones únicas que se dan aquí. Use el algoritmo del disparo lineal para aproximar las soluciones de los siguientes problemas, y después compare los resultados con las soluciones reales.

a. $y'' + y = 0, \quad 0 \leq x \leq \frac{\pi}{4}, \quad y(0) = 1, \quad y\left(\frac{\pi}{4}\right) = 1$; use $h = \frac{\pi}{4}$; solución real $y(x) = \cos x + (\sqrt{2} - 1)\operatorname{sen} x$.
b. $y'' + 4y = \cos x, \quad 0 \leq x \leq \frac{\pi}{4}, \quad y(0) = 0, \quad y(\frac{\pi}{4}) = 0$; use $h = \frac{\pi}{20}$; solución real $y(x) = -\frac{1}{3}\cos 2x - \frac{\sqrt{2}}{6}\operatorname{sen} 2x + \frac{1}{3}\cos x$.
c. $y'' = -\frac{4}{x}y' - \frac{2}{x^2}y + \frac{2}{x^2}\ln x, \quad 1 \leq x \leq 2, \quad y(1) = \frac{1}{2}, \quad y(2) = \ln 2$; use $h = 0.05$; solución real $y(x) = \frac{4}{x} - \frac{2}{x^2} + \ln x - \frac{3}{2}$.
d. $y'' = 2y' - y + xe^x - x, \quad 0 \leq x \leq 2, \quad y(0) = 0, \quad y(2) = -4$; use $h = 0.2$; solución real $y(x) = \frac{1}{6}x^3e^x - \frac{5}{3}xe^x + 2e^x - x - 2$.

5. Use el algoritmo del disparo lineal para aproximar la solución $y = e^{-10x}$ del problema con valor de frontera

$$y'' = 100y, \quad 0 \le x \le 1, \quad y(0) = 1, \quad y(1) = e^{-10}.$$

Use $h = 0.1$ y 0.05.

6. Escriba como sistemas de primer orden los problemas (11.3) y (11.4) de valor inicial de segundo orden, y derive las ecuaciones necesarias para resolver los sistemas mediante el método de Runge-Kutta de cuarto orden para sistemas.

7. Representamos con u el potencial electrostático entre dos esferas metálicas concéntricas de radios R_1 y R_2 ($R_1 < R_2$), tales que el potencial de la esfera interior se mantenga constante en V_1 volts y el potencial de la esfera exterior sea 0 volts. El potencial de la región situada entre ambas esferas está regido por la ecuación de Laplace, que en esta aplicación particular se reduce a

$$\frac{d^2u}{dr^2} + \frac{2}{r}\frac{du}{dr} = 0, \quad R_1 \le r \le R_2, \quad u(R_1) = V_1, \quad u(R_2) = 0.$$

Suponga que $R_1 = 2$ plg, $R_2 = 4$ plg y que $V_1 = 110$ volts.

a. Aproxime $u(3)$ por medio del algoritmo del disparo lineal.

b. Compare los resultados de la parte (a) con el potencial real $u(3)$, donde

$$u(r) = \frac{V_1R_1}{r}\left(\frac{R_2 - r}{R_2 - R_1}\right).$$

8. Demuestre que bajo las hipótesis del corolario 11.2, si y_2 es la solución de $y'' = p(x)y' + q(x)y$ y $y_2(a) = y_2(b) = 0$, entonces $y_2 \equiv 0$.

9. Considere el problema con valor de frontera

$$y'' + y = 0, \quad 0 \le x \le b, \quad y(0) = 0, \quad y(b) = B.$$

Encuentre los valores de b y B para que el problema con valor de frontera

a. No tenga solución;

b. Tenga exactamente una solución;

c. Tenga una infinidad de soluciones.

10. Intente aplicar el ejercicio 9 al problema de valor de frontera

$$y'' - y = 0, \quad 0 \le x \le b, \quad y(0) = 0, \quad y(b) = B.$$

¿Qué sucede? ¿De qué manera ambos problemas se relacionan con el corolario 11.2?

11.2 El método del disparo para problemas no lineales

El método del disparo para el problema no lineal con valor de frontera de segundo orden

$$y'' = f(x, y, y'), \quad a \le x \le b, \quad y(a) = \alpha, \quad y(b) = \beta, \tag{11.6}$$

se parece al caso lineal, excepto que la solución del problema no lineal no puede expresarse como una combinación lineal de las soluciones a los problemas de dos valores iniciales. Necesitamos, en cambio, utilizar las soluciones de una *sucesión* de problemas de

valor inicial de la forma que contengan un parámetro t, para aproximar la solución al problema de valor de frontera.

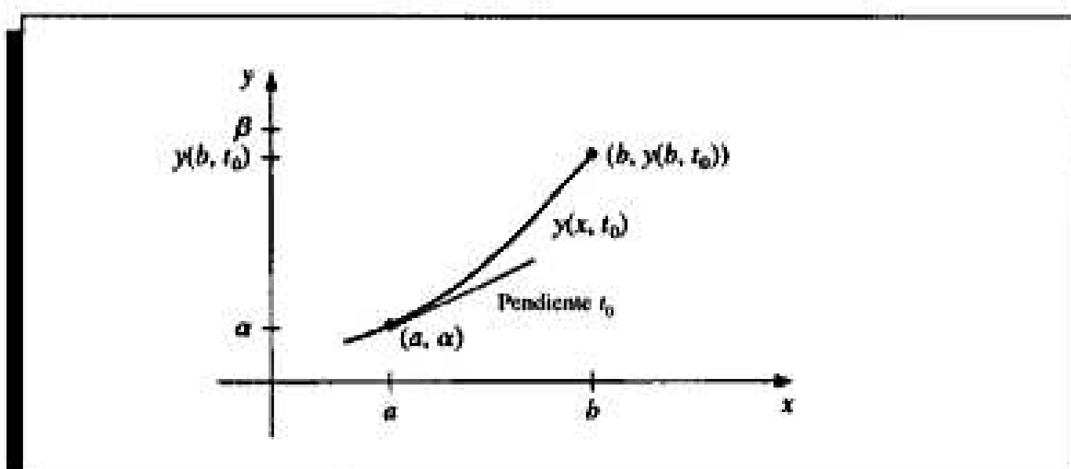
$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y'(a) = t. \tag{11.7}$$

Esto lo hacemos escogiendo los parámetros $t = t_k$ de tal forma que

$$\lim_{k\to\infty} y(b, t_k) = y(b) = \beta,$$

donde $y(x, t_k)$ denota la solución al problema con valor inicial (11.7) con $t = t_k$ y $y(x)$ denota la solución al problema con valor de frontera (11.6).

Figura 11.2



Esta técnica se conoce con el nombre de método "del disparo", por la analogía con el procedimiento de dispararles a objetos situados en un blanco fijo. (Véase la Fig. 11.2.) Comenzamos con un parámetro t_0 que determina la elevación inicial a la cual se le dispara al objetivo desde el punto (a, α) y a lo largo de la curva descrita por la solución al problema de valor inicial:

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y'(a) = t_0.$$

Si $y(b, t_0)$ no está suficientemente cerca de β, corregimos la aproximación seleccionando las elevaciones t_1, t_2, y así sucesivamente, hasta que $y(b, t_k)$ esté bastante cerca de "acertar en el blanco" β. (Véase la Fig. 11.3.)

Para determinar los parámetros t_k, supongamos que un problema de valor de frontera de tipo (11.6) satisface las hipótesis del teorema 11.1. Si $y(x, t)$ denota la solución del problema de valor inicial (11.7), el problema consistirá en determinar t tal que

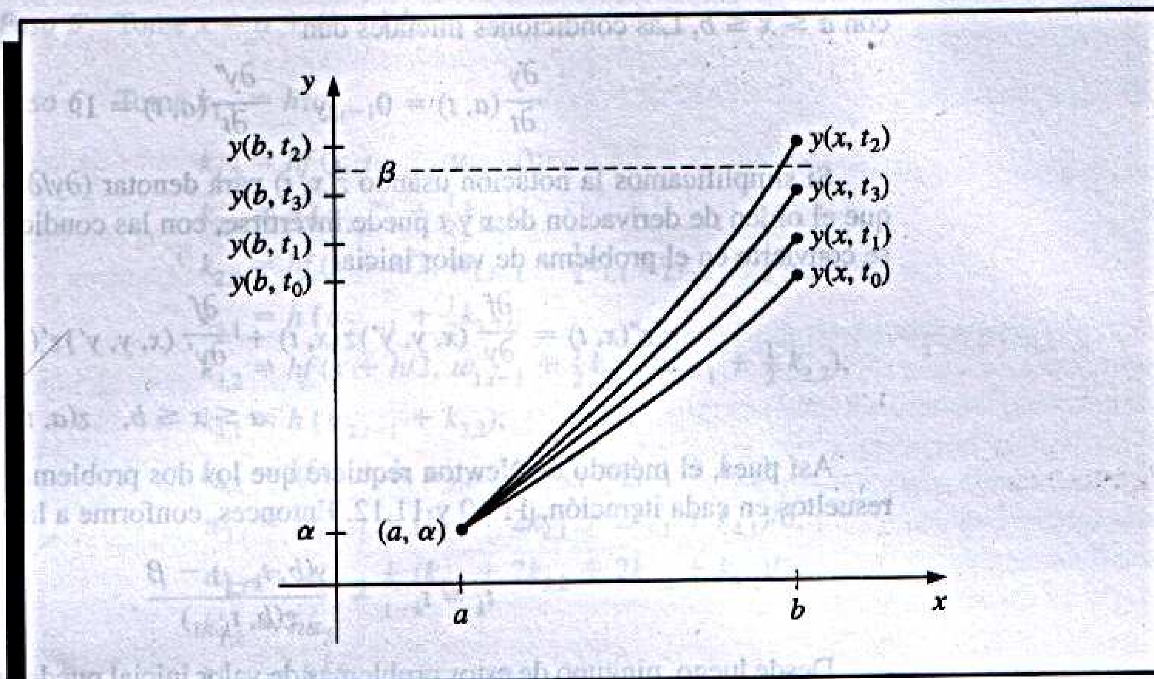
$$y(b, t) - \beta = 0. \tag{11.8}$$

Esta es una ecuación no lineal como las que vimos en el capítulo 2 y, por lo mismo, disponemos de varios métodos.

Si queremos emplear el método de la secante para resolver el problema, necesitamos elegir las aproximaciones iniciales t_0 y t_1 y luego generar los términos restantes de la sucesión mediante

$$t_k = t_{k-1} - \frac{(y(b, t_{k-1}) - \beta)(t_{k-1} - t_{k-2})}{y(b, t_{k-1}) - y(b, t_{k-2})}, \; k = 2, 3, \ldots .$$

Figura 11.3



Para generar la sucesión $\{t_k\}$ con el método de Newton, que es más poderoso, sólo necesitamos una aproximación inicial t_0. Sin embargo, la iteración tiene la forma

$$t_k = t_{k-1} - \frac{y(b, t_{k-1}) - \beta}{\frac{dy}{dt}(b, t_{k-1})} \tag{11.9}$$

y requiere conocer $(dy/dt)(b, t_{k-1})$. Esto presenta un problema porque no se conoce una representación explícita de $y(b, t)$; conocemos sólo los valores $y(b, t_0), y(b, t_1), \ldots, y(b, t_{k-1})$.

Supóngase que reescribimos el problema de valor inicial (11.7), haciendo énfasis en que la solución se basa tanto en x como en el parámetro t:

$$y''(x, t) = f(x, y(x, t), y'(x, t)), \quad a \leq x \leq b, \quad y(a, t) = \alpha, \quad y'(a, t) = t. \tag{11.10}$$

Hemos conservado la notación prima para indicar la derivada respecto a x. Puesto que necesitamos determinar $(dy/dt)(b, t)$ cuando $t = t_{k-1}$, primero tomamos la derivada parcial de (11.10) respecto a t. Esto significa que

$$\begin{aligned}\frac{\partial y''}{\partial t}(x, t) &= \frac{\partial f}{\partial t}(x, y(x, t), y'(x, t))\\ &= \frac{\partial f}{\partial x}(x, y(x, t), y'(x, t))\,\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}(x, y(x, t), y'(x, t)\,\frac{\partial y}{\partial t}(x, t)\\ &\quad + \frac{\partial f}{\partial y'}(x, y(x, t), y'(x, t),\,\frac{\partial y'}{\partial t}(x, t).\end{aligned}$$

Dado que x y t son independientes, $\partial x/\partial t = 0$ y

$$\frac{\partial y''}{\partial t}(x, t) = \frac{\partial f}{\partial y}(x, y(x, t), y'(x, t))\,\frac{\partial y}{\partial t}(x, t) + \frac{\partial f}{\partial y'}(x, y(x, t), y'(x, t))\,\frac{\partial y'}{\partial t}(x, t), \tag{11.11}$$

con $a \le x \le b$. Las condiciones iniciales dan

$$\frac{\partial y}{\partial t}(a, t) = 0 \quad \text{y} \quad \frac{\partial y'}{\partial t}(a, t) = 1.$$

Si simplificamos la notación usando $z(x, t)$ para denotar $(\partial y/\partial t)(x, t)$ y si suponemos que el orden de derivación de x y t puede invertirse, con las condiciones iniciales (11.11) se convierte en el problema de valor inicial

$$z''(x, t) = \frac{\partial f}{\partial y}(x, y, y')z(x, t) + \frac{\partial f}{\partial y'}(x, y, y')z'(x, t), \tag{11.12}$$

$$a \le x \le b, \quad z(a, t) = 0, \quad z'(a, t) = 1.$$

Así pues, el método de Newton requiere que los dos problemas de valor inicial sean resueltos en cada iteración, 11.10 y 11.12. Entonces, conforme a la ecuación (11.9),

$$t_k = t_{k-1} - \frac{y(b, t_{k-1}) - \beta}{z(b, t_{k-1})}. \tag{11.13}$$

Desde luego, ninguno de estos problemas de valor inicial puede resolverse exactamente; las soluciones se aproximan con uno de los métodos expuestos en el capítulo 5. En el algoritmo 11.2 se emplea el método de Runge-Kutta de cuarto orden para aproximar las dos soluciones que requiere el método de Newton. En el ejercicio 4 se considera un procedimiento semejante con el método de la secante.

ALGORITMO 11.2

Disparo no lineal con el método de Newton

Para aproximar la solución del problema no lineal con valor en la frontera

$$y'' = f(x, y, y'), \quad a \le x \le b, \quad y(a) = \alpha, \quad y(b) = \beta:$$

(*Nota: las ecuaciones* 11.10 *y* 11.12 *se escriben y se resuelven como sistemas de primer orden.*)

ENTRADA extremos a, b; condiciones de frontera α, β; número de subintervalos $N \ge 2$; tolerancia TOL; número máximo de iteraciones M.

SALIDA aproximaciones $w_{1,i}$ a $y(x_i)$; $w_{2,i}$ a $y'(x_i)$ para toda $i = 0, 1, \ldots, N$, o bien un mensaje de que se excedió el número máximo de iteraciones.

Paso 1 Tome $h = (b - a)/N$;
$k = 1$;
$TK = (\beta - \alpha)/(b - a)$. (*Nota: TK también puede ser entrada.*)

Paso 2 Mientras ($k \le M$) haga los pasos 3-10.

Paso 3 Tome $w_{1,0} = \alpha$;
$w_{2,0} = TK$;
$u_1 = 0$;
$u_2 = 1$.

Paso 4 Para $i = 1, \ldots, N$ haga los pasos 5 y 6.
(*El método Runge-Kutta para sistemas se utiliza en los pasos 5 y 6.*)

Paso 5 Tome $x = a + (i - 1)h$.

Paso 6 Tome $k_{1,1} = hw_{2,i-1}$;

$$k_{1,2} = hf(x, w_{1,i-1}\ w_{2,i-1});$$
$$k_{2,1} = h\,(w_{2,i-1} + \tfrac{1}{2}k_{1,2});$$
$$k_{2,2} = hf(x + h/2,\ w_{1,i-1} + \tfrac{1}{2}k_{1,1}\ w_{2,i-1} + \tfrac{1}{2}k_{1,2});$$
$$k_{3,1} = h\,(w_{2,i-1} + \tfrac{1}{2}k_{2,2});$$
$$k_{3,2} = hf(x + h/2,\ w_{1,i-1} + \tfrac{1}{2}k_{2,1}\ w_{2,i-1} + \tfrac{1}{2}k_{2,2});$$
$$k_{4,1} = h\,(w_{2,i-1} + k_{3,2});$$
$$k_{4,2} = hf(x + h,\ w_{1,i-1} + k_{3,1}\ w_{2,i-1} + k_{3,2});$$
$$w_{1,i} = w_{1,i-1} + (k_{1,1} + 2k_{2,1} + 2k_{3,1} + k_{4,1})/6;$$
$$w_{2,i} = w_{2,i-1} + (k_{1,2} + 2k_{2,2} + 2k_{3,2} + k_{4,2})/6;$$
$$k'_{1,1} = hu_2;$$
$$k'_{1,2} = h\left[f_y(x,\ w_{1,i-1},\ w_{2,i-1})u_1 + f_{y'}(x,\ w_{1,i-1},\ w_{2,i-1})u_2\right];$$
$$k'_{2,1} = h\left[u_2 + \tfrac{1}{2}\,k'_{1,2}\right];$$
$$k'_{2,2} = h\left[f_y(x + h/2,\ w_{1,i-1},\ w_{2,i-1})\,(u_1 + \tfrac{1}{2}\,k'_{1,1}) + f_{y'}(x + h/2,\ w_{1,i-1},\ w_{2,i-1})\,(u_2 + \tfrac{1}{2}\,k'_{1,2})\right];$$
$$k'_{3,1} = h\left[u_2 + \tfrac{1}{2}\,k'_{2,2}\right];$$
$$k'_{3,2} = h\left[f_y(x + h/2,\ w_{1,i-1},\ w_{2,i-1})\,(u_1 + \tfrac{1}{2}\,k'_{2,1}) + f_{y'}(x + h/2,\ w_{1,i-1},\ w_{2,i-1})\,(u_2 + \tfrac{1}{2}k'_{2,2})\right];$$
$$k'_{4,1} = h\,(u_2 + k'_{3,2});$$
$$k'_{4,2} = h\left[f_y(x + h,\ w_{1,i-1},\ w_{2,i-1})\,(u_1 + k'_{3,1}) + f_{y'}(x + h,\ w_{1,i-1},\ w_{2,i-1})\,(u_2 + k'_{3,2})\right];$$
$$u_1 = u_1 + \tfrac{1}{6}\left[k'_{1,1} + 2\,k'_{2,1} + 2k'_{3,1} + k'_{4,1}\right];$$
$$u_2 = u_2 + \tfrac{1}{6}\left[k'_{1,2} + 2\,k'_{2,2} + 2k'_{3,2} + k'_{4,2}\right].$$

Paso 7 Si $|w_{1,N} - \beta| \leq TOL$, entonces haga los pasos 8 y 9.

Paso 8 Para $i = 0, 1, \ldots, N$
entonces $x = a + ih$;
SALIDA $(x,\ w_{1,i},\ w_{2,i})$.

Paso 9 (*Procedimiento terminado.*)
PARAR.

Paso 10 Tome $TK = TK - \dfrac{w_{1,N} - \beta}{u_1}$;

(*El método de Newton se utiliza para calcular* TK.)
$k = k + 1$.

Paso 11 SALIDA (*'Número máximo de iteraciones excedido'*);
(*Procedimiento terminado sin éxito.*)
PARAR. ■

El valor $t_0 = TK$ escogido en el paso 1 es la pendiente de la recta que pasa por (a, α) y por (b, β). Si el problema satisface las hipótesis del teorema 11.1, cualquier elección de t_0 dará convergencia; pero con una buena elección de t_0, la convergencia mejorará y el procedimiento funcionará en muchos problemas que no satisfacen estas hipótesis.

EJEMPLO 1 Considere el problema con valor de frontera

$$y'' = \frac{1}{8}(32 + 2x^3 - yy'), \quad 1 \leq x \leq 3, \quad y(1) = 17, \quad y(3) = \frac{43}{3},$$

que tiene la solución exacta $y(x) = x^2 + 16/x$.

Si queremos aplicar el método del disparo del algoritmo 11.2 a este problema, hay que aproximar los problemas de valor inicial

$$y'' = \frac{1}{8}(32 + 2x^3 - yy'), \quad 1 \leq x \leq 3, \quad y(1) = 17, \quad y'(1) = t_k,$$

y

$$z'' = \frac{\partial f}{\partial y}z + \frac{\partial f}{\partial y'}z' = -\frac{1}{8}(y'z + yz'), \quad 1 \leq x \leq 3, \quad z(1) = 0, \quad z'(1) = 1,$$

en cada paso de la iteración.

Tabla 11.2

x_i	$w_{1,i}$	$y(x_i)$	$\lvert w_{1,i} - y(x_i)\rvert$
1.0	17.000000	17.000000	
1.1	15.755495	15.755455	4.06×10^{-5}
1.2	14.773389	14.773333	5.60×10^{-5}
1.3	13.997752	13.997692	5.94×10^{-5}
1.4	13.388629	13.388571	5.71×10^{-5}
1.5	12.916719	12.916667	5.23×10^{-5}
1.6	12.560046	12.560000	4.64×10^{-5}
1.7	12.301805	12.301765	4.02×10^{-5}
1.8	12.128923	12.128889	3.14×10^{-5}
1.9	12.031081	12.031053	2.84×10^{-5}
2.0	12.000023	12.000000	2.32×10^{-5}
2.1	12.029066	12.029048	1.84×10^{-5}
2.2	12.112741	12.112727	1.40×10^{-5}
2.3	12.246532	12.246522	1.01×10^{-5}
2.4	12.426673	12.426667	6.68×10^{-6}
2.5	12.650004	12.650000	3.61×10^{-6}
2.6	12.913847	12.913847	9.17×10^{-7}
2.7	13.215924	13.215926	1.43×10^{-6}
2.8	13.554282	13.554286	3.46×10^{-6}
2.9	13.927236	13.927241	5.21×10^{-6}
3.0	14.333327	14.333333	6.69×10^{-6}

Si la técnica de detención o paro requiere

$$|w_{1,N}(t_k) - y(3)| \leq 10^{-5},$$

necesitamos entonces cuatro iteraciones y $t_4 = -14.000203$. Los resultados obtenidos con este valor de t se muestran en la tabla 11.2. ■

Aunque el método de Newton que se usa con las técnicas de disparo requiere la resolución de un problema adicional de valor inicial, en la mayoría de los casos será más rápido que el método de la secante. Ambos métodos convergen sólo localmente, pues requieren buenas aproximaciones iniciales. Al lector que desee una explicación general de la convergencia de los métodos del disparo con problemas no lineales le recomendamos el excelente libro de Keller [Keller, H]. En él se describen las condiciones de frontera más generales. También conviene señalar que el método del disparo para problemas no lineales es vulnerable a los errores de redondeo, especialmente si las soluciones $y(x)$ y $z(x, t)$ son funciones que crecen rápidamente en $[a, b]$.

CONJUNTO DE EJERCICIOS 11.2

1. Use el algoritmo no lineal de disparo con $h = 0.5$ para aproximar la solución al problema con valor de frontera.

$$y'' = -(y')^2 - y + \ln x, \quad 1 \leq x \leq 2, \quad y(1) = 0, \quad y(2) = \ln 2.$$

Compare sus resultados con la solución real $y = \ln x$.

2. Use el algoritmo del disparo no lineal con $h = 0.25$ para aproximar la solución al problema con valor de frontera

$$y'' = 2y^3, \quad -1 \leq x \leq 0, \quad y(-1) = \tfrac{1}{2}, \quad y(0) = \tfrac{1}{3}.$$

Compare los resultados con las soluciones reales $y(x) = 1/(x + 3)$.

3. Aplique el método del disparo no lineal con $TOL = 10^{-4}$ para aproximar la solución a los siguientes problemas con valor de frontera. Se incluye la solución real para facilitar la comparación de los resultados.

   a. $y'' = y^3 - yy'$, $1 \leq x \leq 2$, $y(1) = \frac{1}{2}$, $y(2) = \frac{1}{3}$; use $h = 0.1$ y compare después los resultados con $y(x) = (x + 1)^{-1}$.

   b. $y'' = 2y^3 - 6y - 2x^3$, $1 \leq x \leq 2$, $y(1) = 2$, $y(2) = \frac{5}{2}$; use $h = 0.1$ y compare después los resultados con $y(x) = x + x^{-1}$.

   c. $y'' = y' + 2(y - \ln x)^3 - x^{-1}$, $2 \leq x \leq 3$, $y(2) = \frac{1}{2}, + \ln 2$, $y(3) = \frac{1}{3} = + \ln 3$; use $h = 0.1$ y compare después los resultados con $y(x) = x^{-1} + \ln x$.

   d. $y'' = [x^2 (y')^2 - 9y^2 + 4x^6]/x^5$, $1 \leq x \leq 2$, $y(1) = 0$, $y(2) = \ln 256$; use $h = 0.05$ y compare después los resultados con $y(x) = x^3 \ln x$.

4. Cambie el algoritmo 11.2 para incorporar el método de la secante en vez del método de Newton. Utilice $t_0 = (\beta - \alpha)/(b - a)$ y $t_1 = t_0 + (\beta - y(b, t_0))/(b - a)$.

5. Repita el ejercicio 3(a) y 3(c) usando el algoritmo del método de la secante derivado en el ejercicio 4; después compare la cantidad de iteraciones que se requieren en ambos métodos.

6. La ecuación de Van der Pol,

$$y'' - \mu(y^2 - 1)y' + y = 0, \quad \mu > 0,$$

rige el flujo de la corriente en un tubo al vacío con tres elementos internos. Sea $\mu = \frac{1}{2}$, $y(0) = 0$, y $y(2) = 1$. Aproxime la solución $y(t)$ para $t = 0.2i$, donde $1 \leq i \leq 9$.

11.3 Métodos de diferencias finitas para los problemas lineales

Aunque los métodos del disparo pueden emplearse en los problemas lineales y no lineales de valor de frontera, a menudo presentan problemas de inestabilidad. Los métodos que expondremos en esta sección tiene mejores características de estabilidad, pero generalmente hay que trabajar más para obtener la exactitud especificada.

Los métodos para la resolución de problemas de valor de frontera que contienen diferencias finitas reemplazan las derivadas en la ecuación diferencial mediante una aproximación de cociente de diferencias adecuada, como la considerada en la sección 4.1. Se selecciona el cociente de diferencias para mantener un orden especificado del error de truncamiento. Pero, por la inestabilidad de las aproximaciones de diferencias finitas a las derivadas, no podemos escoger un parámetro h demasiado pequeño.

El método de diferencias finitas para el problema de valor de frontera de segundo orden,

$$y'' = p(x)y' + q(x)y + r(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta, \tag{11.14}$$

requiere utilizar las aproximaciones del cociente de diferencias para aproximar tanto a y' como a y''. Primero, seleccionamos un entero $N > 0$ y dividimos el intervalo $[a, b]$ en $(N + 1)$ subintervalos iguales cuyos extremos son los puntos de malla $x_i = a + ih$, para $i = 0, 1, \ldots, N + 1$, donde $h = (b - a)/(N + 1)$. Al escoger h de este modo, se facilita la aplicación de un algoritmo matricial explicado en el capítulo 6, con el cual se resuelve un sistema lineal que contenga una matriz de $N \times N$.

En los puntos de red del interior, x_i, para $i = 1, 2, \ldots, N$, la ecuación diferencial a aproximar es

$$y''(x_i) = p(x_i)\, y'(x_i) + q(x_i)y(x_i) + r(x_i). \tag{11.15}$$

Al desarrollar y en el tercer polinomio de Taylor alrededor de x_i evaluada en x_{i+1} y x_{i-1}, tenemos, suponiendo que $y \in C^4\,[x_{i-1}, x_{i+1}]$,

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + hy'(x_i) + \frac{h^2}{2}\, y''(x_i) + \frac{h^3}{6}\, y'''\,(x_i) + \frac{h^4}{24}\, y^{(4)}\,(\xi_i^+),$$

para alguna ξ_i^+ en (x_i, x_{i+1}), y

$$y(x_{i-1}) = y(x_i - h) = y(x_i) - hy'(x_i) + \frac{h^2}{2}\, y''(x_i) - \frac{h^3}{6}\, y'''(x_i) + \frac{h^4}{24}\, y^{(4)}\,(\xi_i^-),$$

para alguna ξ_i^- en (x_{i-1}, x_i), si se suman estas ecuaciones, tenemos

$$y(x_{i+1}) + y(x_{i-1}) = 2y(x_i) + h^2y''(x_i) + \frac{h^4}{24}[y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-)],$$

y al despejar $y''(x_i)$ se obtiene

$$y''(x_i) + \frac{1}{h^2}[y(x_{i+1}) - 2y(x_i) + y(x_{i-1})] - \frac{h^2}{24}[y^{(4)}(\xi_i^+) + y^{(4)}(\xi_i^-)].$$

Podemos aplicar el teorema del valor intermedio para simplificar aún más esta expresión y transformarla en

$$y''(x_i) + \frac{1}{h^2}[y(x_{i+1}) - 2y(x_i) + y(x_{i-1})] - \frac{h^2}{12}y^{(4)}(\xi_i) \tag{11.16}$$

para alguna ξ_i en (x_{i-1}, x_{i+1}). A esto se le llama **fórmula de las diferencias centradas** para $y''(x_i)$.

De manera semejante se obtiene una fórmula de este tipo para $y'(x_i)$ (los detalles se dan en la sección 4.1), que da por resultado

$$y'(x_i) = \frac{1}{2h}[y(x_{i+1}) - y(x_{i-1})] - \frac{h^2}{6}y'''(\eta_i) \tag{11.17}$$

para alguna η_i en (x_{i-1}, x_{i+1}).

La utilización de las fórmulas de diferencias centradas en la ecuación (11.15) genera la ecuación

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = p(x_i)\left[\frac{y(x_{i+1}) - y(x_{i-1})}{2h}\right] + q(x_i)y(x_i)$$
$$+ r(x_i) - \frac{h^2}{12}\left[2p(x_i)y'''(\eta_i) - y^{(4)}(\xi_i)\right].$$

Un método de diferencias finitas con error de truncamiento de orden $O(h^2)$ se obtiene empleando esta ecuación junto con las condiciones de frontera $y(a) = \alpha$ y $y(b) = \beta$ para definir

$$w_0 = \alpha, \quad w_{N+1} = \beta$$

y

$$\left(\frac{-w_{i+1} + 2w_i - w_{i-1}}{h^2}\right) + p(x_i)\left(\frac{w_{i+1} - w_{i-1}}{2h}\right) + q(x_i)w_i = -r(x_i), \tag{11.18}$$

para toda $i = 1, 2, \ldots, N$.

En la forma que consideraremos, la ecuación (11.18) se reescribe como

$$-\left(1 + \frac{h}{2}p(x_i)\right)w_{i-1} + \left(2 + h^2q(x_i)\right)w_i - \left(1 - \frac{h}{2}p(x_i)\right)w_{i+1} = -h^2r(x_i),$$

y el sistema de ecuaciones resultante se expresa en forma de la matriz tridiagonal de $N \times N$

$$A\mathbf{w} = \mathbf{b}, \quad \text{donde} \tag{11.19}$$

$$A = \begin{bmatrix} 2 + h^2q(x_1) & -1 + \frac{h}{2}p(x_1) & 0 & \cdots & 0 \\ -1 - \frac{h}{2}p(x_2) & 2 + h^2q(x_2) & -1 + \frac{h}{2}p(x_2) & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 + \frac{h}{2}p(x_{N-1}) \\ 0 & \cdots & 0 & -1 - \frac{h}{2}p(x_N) & 2 + h^2q(x_N) \end{bmatrix},$$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N-1} \\ w_N \end{bmatrix} \quad \text{y} \quad \mathbf{b} = \begin{bmatrix} -h^2r(x_1) + \left(1 + \frac{h}{2}p(x_1)\right)w_0 \\ -h^2r(x_2) \\ \vdots \\ -h^2r(x_{N-1}) \\ -h^2r(x_N) + \left(1 + \frac{h}{2}p(x_N)\right)w_{N+1} \end{bmatrix}.$$

El siguiente teorema establece las condiciones bajo las cuales el sistema lineal (11.19) tiene una solución única. Su demostración es consecuencia del teorema 6.29 y se considera en el ejercicio 9.

Teorema 11.3 Supongamos que p, q y r son continuas en $[a, b]$. Si $q(x) \geq 0$ en $[a, b]$, entonces el sistema lineal tridiagonal (11,19) tiene una solución única siempre y cuando $h < 2/L$, donde $L = \max_{a \leq x \leq b} |p(x)|$. ■

Conviene señalar que las hipótesis del teorema 11.3 garantizan una solución única al problema de valor de frontera (11.14), pero no que $y \in C^4[a, b]$. Para asegurarnos de que el error de truncamiento tiene el orden $O(h^2)$, debemos establecer que $y^{(4)}$ es continua en $[a, b]$.

En el algoritmo 11.3 se ejecuta el método lineal de diferencias finitas.

ALGORITMO 11.3

Método lineal de diferencias finitas

Para aproximar la solución al problema de valor de frontera

$$y'' = p(x)y' + q(x)y + r(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta:$$

ENTRADA extremos a, b; condiciones de frontera α, β; entero $N \geq 2$.

SALIDA aproximaciones w_i a $y(x_i)$ para toda $i = 0, 1, \ldots, N + 1$.

Paso 1 Tome $h = (b - a)/(N + 1)$;
$x = a + h$;
$a_1 = 2 + h^2q(x)$;
$b_1 = -1 + (h/2)p(x)$;
$d_1 = -h^2r(x) + (1 + (h/2)p(x))\alpha$.

Paso 2 Para $i = 2, \ldots, N-1$
tome $x = a + ih$;
$a_i = 2 + h^2 q(x)$;
$b_i = -1 + (h/2)p(x)$;
$c_i = -1 - (h/2)p(x)$;
$d_i = -h^2 r(x)$.

Paso 3 Tome $x = b - h$;
$a_N = 2 + h^2 q(x)$;
$c_N = -1 - (h/2)p(x)$;
$d_N = -h^2 r(x) + (1 - (h/2)p(x))\beta$.

Paso 4 Tome $l_1 = a_1$; ***(Los pasos 4-8 resuelven un sistema lineal tridiagonal utilizando el algoritmo 6.7.)***
$u_1 = b_1/a_1$;
$z_1 = d_1/l_1$.

Paso 5 Para $i = 2, \ldots, N-1$, tome $l_i = a_i - c_i u_{i-1}$;
$u_i = b_i/l_i$;
$z_i = (d_i - c_i z_{i-1})/l_i$.

Paso 6 Tome $l_N = a_N - c_N u_{N-1}$;
$z_N = (d_N - c_N z_{N-1})/l_N$.

Paso 7 Tome $w_0 = \alpha$;
$w_{N+1} = \beta$.
$w_N = z_N$.

Paso 8 Para $i = N-1, \ldots, 1$, tome $w_i = z_i - u_i w_{i+1}$.

Paso 9 Para $i = 0, \ldots, N+1$, tome $x = a + ih$;
SALIDA (x, w_i).

Paso 10 PARAR. (*Procedimiento terminado.*) ■

EJEMPLO 1 Utilizaremos el algoritmo 11.3 para aproximar la solución al problema lineal con valor de frontera

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad 1 \le x \le 2, \quad y(1) = 1, \quad y(2) = 2,$$

que también se aproxima mediante el método del disparo considerado en el ejemplo 2 de la sección 11.1. En este ejemplo usaremos $N = 9$, de modo que $h = 0.1$ y tenemos el mismo espaciado que en el ejemplo 2 de la sección 11.1. Los resultados se proporcionan en la tabla 11.3.

Nótese que estos resultados son mucho menos exactos que los obtenidos en el ejemplo 2 de la sección 11.1. Esto se debe a que el método empleado en ese ejemplo incluía un método de Runge-Kutta con un error de truncamiento de orden $O(h^4)$, mientras que el método de diferencias usado aquí presenta un error del orden $O(h^2)$. ■

Si queremos obtener un método de diferencias con mayor precisión, podemos recurrir a varias opciones. Si usamos la serie de Taylor de quinto orden para aproximar $y''(x_i)$ y $y'(x_i)$ se producirá un error de truncamiento que involucra al término h^4. Sin embargo, este proceso requiere emplear múltiplos no sólo de $y(x_i + 1)$ y $y(x_i - 1)$, sino también de

Tabla 11.3

x_i	w_i	$y(x_i)$	$\|w_i - y(x_i)\|$
1.0	1.00000000	1.00000000	
1.1	1.09260052	1.09262930	2.88×10^{-5}
1.2	1.18704313	1.18708484	4.17×10^{-5}
1.3	1.28333687	1.28338236	4.55×10^{-5}
1.4	1.38140205	1.38144595	4.39×10^{-5}
1.5	1.48112026	1.48115942	3.92×10^{-5}
1.6	1.58235990	1.58239246	3.26×10^{-5}
1.7	1.68498902	1.68501396	2.49×10^{-5}
1.8	1.78888175	1.78889853	1.68×10^{-5}
1.9	1.89392110	1.89392951	8.41×10^{-6}
2.0	2.00000000	2.00000000	

$y(x_{i+1})$ y $y(x_{i-1})$ en las fórmulas de aproximación para $y''(x_i)$ y $y'(x_i)$. Esto da origen a tener problemas en $i = 0$ y en $i = N$. Más aún, el sistema resultante de ecuaciones análogas a (11.19) no presenta la forma tridiagonal, y su solución requiere una labor de cálculo mucho mayor.

En vez de intentar obtener en esta forma un método de diferencias con un error de truncamiento de orden superior, generalmente resulta mejor considerar una reducción del tamaño de paso. Además, es posible demostrar (véase, por ejemplo, [Keller, H, p. 81]) que la extrapolación de Richardson puede emplearse eficazmente con este método, ya que el término de error se expresa en potencias pares de h con coeficientes independientes de h, siempre y cuando y sea suficientemente diferenciable.

EJEMPLO 2 Los resultados de la tabla 11.4 se obtienen al aplicar la extrapolación de Richardson para aproximar la solución al problema con valor de frontera

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\operatorname{sen}(\ln x)}{x^2}, \quad 1 \le x \le 2, \quad y(1) = 1, \quad y(2) = 2,$$

con $h = 0.1$, 0.05 y 0.025. La primera extrapolación es

Tabla 11.4

x_i	$w_i(h = 0.1)$	$w_i(h = 0.05)$	$w_i(h = 0.025)$	Ext_{1i}	Ext_{2i}	Ext_{3i}
1.0	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
1.1	1.09260052	1.09262207	1.09262749	1.09262925	1.09262930	1.09262930
1.2	1.18704313	1.18707436	1.18708222	1.18708477	1.18708484	1.18708484
1.3	1.28333687	1.28337094	1.28337950	1.28338230	1.28338236	1.28338236
1.4	1.38140205	1.38143493	1.38144319	1.38144589	1.38144595	1.38144595
1.5	1.48112026	1.48114959	1.48115696	1.48115937	1.48115941	1.48115942
1.6	1.58235990	1.58238429	1.58239042	1.58239242	1.58239246	1.58239246
1.7	1.68498902	1.68500770	1.68501240	1.68501393	1.68501396	1.68501396
1.8	1.78888175	1.78889432	1.78889748	1.78889852	1.78889853	1.78889853
1.9	1.89392110	1.89392740	1.89392898	1.89392950	1.89392951	1.89392951
2.0	2.00000000	2.00000000	2.00000000	2.00000000	2.00000000	2.00000000

$$\text{Ext}_{1i} = \frac{4w_i(h = 0.05) - w_i(h = 0.1)}{3};$$

la segunda extrapolación es

$$\text{Ext}_{2i} = \frac{4w_i(h = 0.025) - w_i(h = 0.05)}{3};$$

y la última extrapolación es

$$\text{Ext}_{3i} = \frac{16\text{Ext}_{2i} - \text{Ext}_{1i}}{15}.$$

Todos los resultados de Ext_{3i} son correctos con la exactitud de las cifras decimales señaladas. De hecho, si se conservan suficientes dígitos, esta aproximación proporciona resultados que concuerdan con la solución exacta con un error máximo de 6.3×10^{-11} en los puntos de malla, una mejora notable. ■

CONJUNTO DE EJERCICIOS 11.3

1. El problema con valor de frontera

$$y'' = 4(y - x), \quad 0 \le x \le 1, \quad y(0) = 0, \quad y(1) = 2$$

tiene la solución $y(x) = e^2(e^4 - 1)^{-1}(e^{2x} - e^{-2x}) + x$. Use el algoritmo lineal de diferencias finitas para aproximar la solución, y después compare los resultados con la solución real.

   a. Con $h = \frac{1}{2}$; b. Con $h = \frac{1}{4}$.
   c. Extrapole para aproximar $y(1/2)$.

2. El problema con valor de frontera

$$y'' = y' + 2y + \cos x, \quad 0 \le x \le \frac{\pi}{2}, \quad y(0) = -0.3, \quad y\left(\frac{\pi}{2}\right) = -0.1$$

tiene la solución $y(x) = -\frac{1}{10}(\operatorname{sen} x + 3\cos x)$. Use el algoritmo lineal de diferencias finitas para aproximar la solución, y después compare los resultados con la solución real.

   a. Con $h = \frac{\pi}{4}$; b. Con $h = \frac{\pi}{8}$.
   c. Extrapole para aproximar $y(\pi/4)$.

3. Use el algoritmo lineal de diferencias finitas para aproximar la solución a los siguientes problemas con valor de frontera.
   a. $y'' = -3y' + 2y + 2x + 3, \quad 0 \le x \le 1, \; y(0) = 2, \; y(1) = 1$; use $h = 0.1$.
   b. $y'' = -\frac{4}{x}y' + \frac{2}{x^2}y - \frac{2}{x^2}\ln x, \quad 1 \le x \le 2, \; y(1) = -\frac{1}{2}, \; y(2) = \ln 2$; use $h = 0.05$.
   c. $y'' = -(x + 1)y' + 2y + (1 - x^2)e^{-x}, \quad 0 \le x \le 1, \; y(0) = -1, \; y(1) = 0$; use $h = 0.1$.
   d. $y'' = \frac{y'}{x} + \frac{3}{x^2}y + \frac{\ln x}{x} - 1, \quad 1 \le x \le 2, \; y(1) = y(2) = 0$; use $h = 0.1$.

4. Aunque $q(x) < 0$ en los siguientes problemas con valor de frontera, existen soluciones únicas y se dan. Use el algoritmo lineal de diferencias finitas para aproximar las soluciones, y después compare los resultados con las soluciones reales.
   a. $y'' + y = 0, \quad 0 \le x \le \frac{\pi}{4}, \; y(0) = 1, \; y(\frac{\pi}{4}) = 1$; use $h = \frac{\pi}{20}$; solución real $y(x) = \cos x + (\sqrt{2} - 1)\operatorname{sen} x$.

b. $y'' + 4y = \cos x$, $0 \le x \le \frac{\pi}{4}$, $y(0) = 0$, $y(\frac{\pi}{4}) = 0$; use $h = \frac{\pi}{20}$; solución real $y(x) = -\frac{1}{3}\cos 2x - \frac{\sqrt{2}}{6}\operatorname{sen} 2x + \frac{1}{3}\cos x$.

c. $y'' = -\frac{4}{x}y' - \frac{2}{x^2}y + \frac{2\ln x}{x^2}$, $1 \le x \le 2$, $y(1) = \frac{1}{2}$, $y(2) = \ln 2$; use $h = 0.05$; solución real $y(x) = \frac{4}{x} - \frac{2}{x^2} + \ln x - \frac{3}{2}$.

d. $y'' = 2y' - y + xe^x - x$, $0 \le x \le 2$, $y(0) = 0$, $y(2) = -4$; use $h = 0.2$; solución real $y(x) = \frac{1}{6}x^3e^x - \frac{5}{3}xe^x + 2e^x - x - 2$.

5. Use el algoritmo lineal de diferencias finitas para aproximar las soluciones de $y = e^{-10x}$ al problema con valor de frontera

$$y'' = 100y, \quad 0 \le x \le 1, \quad y(0) = 1, \quad y(1) = e^{-10}.$$

Use $h = 0.1$ y 0.05. ¿Puede explicar las consecuencias?

6. Repita el ejercicio 3(a) y (b) empleando la extrapolación explicada en el ejemplo 2.

7. En el primer ejemplo del capítulo se consideró la deflexión de una viga con los extremos soportados sujetos a una carga uniforme. El problema con valor de frontera que rige esta situación física es

$$\frac{d^2w}{dx^2} = \frac{S}{EI}w + \frac{qx}{2EI}(x - l), \quad 0 < x < l,$$

con las condiciones de frontera $w(0) = 0$ y $w(l) = 0$.

Suponga que la viga es de acero y del tipo W10, con las siguientes características: longitud $l = 120$ plg, intensidad de la carga uniforme $q = 100$ lb/pie, módulo de elasticidad $E = 3.0 \times 10^7$ lb/plg^2, esfuerzo en los extremos $S = 1000$ lb y momento central de inercia $I = 625$ plg^4.

a. Aproxime la deflexión $w(x)$ de la viga cada 6 plg.

b. La relación real está dada por

$$w(x) = c_1e^{ax} + c_2e^{-ax} + b(x - l)x + c,$$

donde $c_1 = 7.7042537 \times 10^4$, $c_2 = 7.9207462 \times 10^4$, $a = 2.3094010 \times 10^{-4}$, $b = -4.1666666 \times 10^{-3}$ y $c = -1.5625 \times 10^5$. ¿No rebasa 0.2 plg el error máximo en el intervalo?

c. La ley estatal de la construcción estipula que $\max_{0<x<l} w(x) < 1/300$. ¿Cumple esta viga con el código estatal?

8. La deflexión de una placa rectangular larga y uniformemente cargada, y que se encuentra bajo una fuerza de tensión axial, se rige por la ecuación diferencial de segundo orden. Sea S la fuerza axial y q la intensidad de la carga uniforme. La deflexión W a lo largo de la longitud elemental está dada por

$$W''(x) - \frac{S}{D}W(x) = \frac{-ql}{2D}x + \frac{q}{2D}x^2, \quad 0 \le x \le l, \; W(0) = W(l) = 0,$$

donde l es la longitud de la placa y D es la rigidez de deflexión de la placa. Sean $q = 200$ lb/plg^2, $S = 100$ lb/plg, $D = 8.8 \times 10^7$ lb/plg y $l = 50$ plg. Aproxime la deflexión en intervalos de 1 plg.

9. Demuestre el teorema 11.3. [*Sugerencia:* para utilizar el teorema 6.29, antes demuestre que $|\frac{h}{2}p(x_i)| < 1$ implica que $|-1 - \frac{h}{2}p(x_i)| + |-1 + \frac{h}{2}p(x_i)| = 2$.]

10. Demuestre que, si $y \in C^6[a, b]$ y si $w_0, w_1, \ldots, w_{N+1}$ satisfacen la ecuación 11.18, entonces

$$w_i - y(x_i) = Ah^2 + O(h^4),$$

donde A es independiente de h, a condición de que $q(x) \ge w > 0$ en $[a, b]$ para alguna w.

11.4 Métodos de diferencias finitas para problemas no lineales

Para el caso del problema no lineal general con valor de frontera

$$y'' = f(x, y, y'), \quad a \le x \le b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

el método de diferencias se parece al que se aplicó a los problemas lineales en la sección 11.3. Sin embargo, aquí el sistema de ecuaciones no será lineal y, por lo mismo, se requiere un proceso iterativo para resolverlo.

Para el desarrollo del procedimiento supondremos que f satisface las siguientes condiciones:

1. f y las derivadas parciales f_y y $f_{y'}$ son continuas en

$$D = \{(x, y, y') \mid a \le x \le b, \quad -\infty < y < \infty, -\infty < y' < \infty\};$$

2. $f_y(x, y, y') \ge \delta$ en D, para alguna $\delta > 0$;

3. Existen las constantes k y L, con

$$k = \max_{(x,y,y')\in D} |f_y(x, y, y')| \quad \text{y} \quad L = \max_{(x,y,y')\in D} |f_{y'}(x, y, y')|.$$

Esto garantiza que, conforme al teorema 11.1, exista una solución única.

Al igual que en el caso de la ecuación lineal, dividimos $[a, b]$ en $(N + 1)$ subintervalos iguales cuyos extremos se encuentran en $x_i = a + ih$, para $i = 0, 1, \ldots, N + 1$. Suponer que la solución exacta tiene una cuarta derivada acotada nos permite reemplazar $y''(x_i)$ y $y'(x_i)$ en cada una de las ecuaciones

$$y''(x_i) = f(x_i, y(x_i), y'(x_i))$$

por la fórmula adecuada de diferencias centradas que se incluyó en las ecuaciones 11.16 y 11.17. Esto nos da, para toda $i = 1, 2, \ldots, N$,

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = f\left(x_i, y(x_i), \frac{y(x_{i+1}) - y(x_{i-1})}{2h} - \frac{h^2}{6} y'''(\eta_i)\right) + \frac{h^2}{12} y^{(4)}(\xi_i),$$

para alguna ξ_i y η_i en el intervalo (x_{i-1}, x_{i+1}).

Como en el caso de la ecuación lineal, los resultados del método de diferencias se emplean cuando se eliminan los términos de error y las condiciones de frontera:

$$w_0 = \alpha, \; w_{N+1} = \beta,$$

y

$$-\frac{w_{i+1} - 2w_i + w_{i-1}}{h^2} + f\left(x_i, w_i, \frac{w_{i+1} - w_{i-1}}{2h}\right) = 0,$$

para toda $i = 1, 2, \ldots, N$.

El sistema no lineal de $N \times N$ obtenido con este método,

$$
\begin{aligned}
2w_1 - w_2 + h^2 f\left(x_1, w_1, \frac{w_2 - \alpha}{2h}\right) - \alpha &= 0, \\
-w_1 + 2w_2 - w_3 + h^2 f\left(x_2, w_2, \frac{w_3 - w_1}{2h}\right) &= 0, \\
&\vdots \\
-w_{N-2} + 2w_{N-1} - w_N + h^2 f\left(x_{N-1}, w_{N-1}, \frac{w_N - w_{N-2}}{2h}\right) &= 0, \\
-w_{N-1} + 2w_N + h^2 f\left(x_N, w_N, \frac{\beta - w_{N-1}}{2h}\right) - \beta &= 0
\end{aligned}
\tag{11.20}
$$

tiene una solución única siempre y cuando $h < 2/L$ como se demuestra en [Keller, H, p. 86].

Aplicamos el método de Newton para sistemas no lineales, expuesto en la sección 10.2, para aproximar la solución de este sistema. Se genera una sucesión de iteraciones $\{(w_1^{(k)}, w_2^{(k)}, \ldots, w_N^{(k)})^t\}$ que converge a la solución del sistema (11.20), a condición de que la aproximación inicial $(w_1^{(0)}, w_2^{(0)}, \ldots, w_N^{(0)})^t$ se acerque lo suficiente a la solución $(w_1, w_2, \ldots, w_N)^t$, y de que la matriz jacobiana del sistema no sea singular. En el caso del sistema (11.20), la matriz jacobiana $J(w_1, \ldots, w_N)$ es tridiagonal con el ij-ésimo elemento,

$$
J(w_1, \ldots, w_N)_{ij} = \begin{cases} -1 + \dfrac{h}{2} f_{y'}\left(x_i, w_i, \dfrac{w_{i+1} - w_{i-1}}{2h}\right), & \text{para } i = j - 1 \text{ y } j = 2, \ldots, N, \\ 2 + h^2 f_y\left(x_i, w_i, \dfrac{w_{i+1} - w_{i-1}}{2h}\right), & \text{para } i = j \text{ y } j = 1, \ldots, N, \\ -1 - \dfrac{h}{2} f_{y'}\left(x_i, w_i, \dfrac{w_{i+1} - w_{i-1}}{2h}\right), & \text{para } i = j + 1 \text{ y } j = 1, \ldots, N-1, \end{cases}
$$

donde $w_0 = \alpha$ y $w_{N+1} = \beta$.

El método de Newton para los sistemas no lineales requiere que en cada iteración del sistema lineal de $N \times N$

$$
\begin{aligned}
J(w_1, \ldots, w_N)(v_1, \ldots, v_N)^t
&= -\left(2w_1 - w_2 - \alpha + h^2 f\left(x_1, w_1, \frac{w_2 - \alpha}{2h}\right),\right. \\
&\quad - w_1 + 2w_2 - w_3 + h^2 f\left(x_2, w_2, \frac{w_3 - w_1}{2h}\right), \ldots, \\
&\quad - w_{N-2} + 2w_{N-1} - w_N + h^2 f\left(x_{N-1}, w_{N-1}, \frac{w_N - w_{N-2}}{2h}\right), \\
&\quad \left. - w_{N-1} + 2w_N + h^2 f\left(x_N, w_N, \frac{\beta - w_{N-1}}{2h}\right) - \beta\right)^t
\end{aligned}
$$

se despejen $v_1, v_2, \ldots, v_N$, porque

$$
w_i^{(k)} = w_i^{(k-1)} + v_i, \quad \text{para cada } i = 1, 2, \ldots, N.
$$

Puesto que J es tridiagonal, este no representa un problema tan difícil como podría parecer a primera vista. Podemos aplicar el algoritmo de factorización de Crout para los sistemas tridiagonales 6.7. El proceso se describe detalladamente en el algoritmo 11.4.

ALGORITMO 11.4

Método de las diferencias finitas para problemas no lineales

Para aproximar la solución del problema no lineal de valor de frontera

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta:$$

ENTRADA extremos a, b; condiciones de frontera α, β; entero $N \geq 2$; tolerancia TOL; número máximo de iteraciones M.

SALIDA aproximaciones w_i a $y(x_i)$ para toda $i = 0, 1, \ldots, N + 1$ o un mensaje de que se excedió el número máximo de iteraciones.

Paso 1 Tome $h = (b - a)/(N + 1)$;
$w_0 = \alpha$;
$w_{N+1} = \beta$.

Paso 2 Para $i = 1, \ldots, N$ tome $w_i = \alpha + i\left(\dfrac{\beta - \alpha}{b - a}\right)h$.

Paso 3 Tome $k = 1$.

Paso 4 Mientras $k \leq M$, haga los pasos 5–16.

Paso 5 Tome $x = a + h$;
$t = (w_2 - \alpha)/(2h)$;
$a_1 = 2 + h^2 f_y(x, w_1, t)$;
$b_1 = -1 + (h/2) f_{y'}(x, w_1, t)$;
$d_1 = -\left(2w_1 - w_2 - \alpha + h^2 f(x, w_1, t)\right)$.

Paso 6 Para $i = 2, \ldots, N - 1$
tome $x = a + ih$;
$t = (w_{i+1} - w_{i-1})/(2h)$;
$a_i = 2 + h^2 f_y(x, w_i, t)$;
$b_i = -1 + (h/2) f_{y'}(x, w_i, t)$;
$c_i = -1 - (h/2) f_{y'}(x, w_i, t)$;
$d_i = -\left(2w_i - w_{i+1} - w_{i-1} + h^2 f(x, w_i, t)\right)$.

Paso 7 Tome $x = b - h$;
$t = (\beta - W_{N-1})/(2h)$;
$a_N = 2 + h^2 f_y(x, w_N, t)$;
$c_N = -1 - (h/2) f_{y'}(x, w_N, t)$;
$d_N = -\left(2w_N - w_{N-1} - \beta + h^2 f(x, w_N, t)\right)$.

Paso 8 Tome $l_1 = a_1$; (*Los pasos 8-12 resuelven un sistema lineal tridiagonal utilizando el algoritmo 6.7.*)
$u_1 = b_1/a_1$;
$z_1 = d_1/l_1$.

Paso 9 Para $i = 2, \ldots, N - 1$, tome $l_i = a_i - c_i\mu_{i-1}$;
$u_i = b_i/l_i$;
$z_i = (d_i - c_i z_{i-1})/l_i$.

Paso 10 Tome $l_N = a_N - c_N u_{N-1}$;
$z_N = (d_N - c_N z_{N-1})/l_N$.

Paso 11 Tome $v_N = z_N$;
$w_N = w_N + v_N$.

Paso12 Para $i = N - 1, \ldots, 1$, tome $v_i = z_i - u_i v_{i+1}$;
$w_i = w_i + v_i$.

Paso 13 Si $\|\mathbf{v}\| \leq TOL$, entonces haga los pasos 14 y 15.

Paso 14 Para $i = 0, \ldots, N + 1$, tome $x = a + ih$;
SALIDA (x, w_i).

Paso 15 PARAR. (*Procedimiento terminado con éxito.*)

Paso 16 Tome $k = k + 1$.

Paso 17 SALIDA ('Número máximo de iteraciones excedido');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Puede demostrarse (véase [IK, p. 433]) que este método no lineal de diferencias finitas es del orden $O(h^2)$.

Cuando no es posible verificar el cumplimiento de las condiciones (1), (2) y (3) dadas al inicio de esta exposición, se requiere una buena aproximación inicial, por lo que conviene especificar una cota superior de k y, en caso de que se rebase, una nueva aproximación inicial o una reducción del tamaño de paso considerado. Las aproximaciones iniciales $w_i^{(0)}$ de w_i, para toda $i = 1, 2, \ldots, N$, se obtienen en el paso 2 cruzando una recta a través de (a, α) y de (b, β) y evaluando en x_i.

EJEMPLO 1 Al aplicar el algoritmo 11.4, con $h = 0.1$, al problema no lineal de valor de frontera

$$y'' = \frac{1}{8}(32 + 2x^3 - yy'), \quad 1 \leq x \leq 3, \quad y(1) = 17, \quad y(3) = \frac{43}{3},$$

se obtienen los resultados mostrados en la tabla 11.5. El procedimiento de paro o detención utilizado en este ejemplo consistía en iterar hasta que los valores de las iteraciones sucesivas difirieran en menos de 10^{-8}, lo cual se logró con cuatro iteraciones. Nótese que el problema de este ejemplo es el mismo que se consideró en el método del disparo no lineal, en el ejemplo 1 de la sección 11.2. ■

También podemos usar el procedimiento de extrapolación de Richardson con el método no lineal de diferencias finitas. En la tabla 11.6 se anotan los resultados obtenidos cuando aplicamos este método a nuestro ejemplo con $h = 0.1$, 0.05 y 0.025, con cuatro iteraciones en cada caso. La notación es la misma que la del ejemplo 2 de la sección 11.3 y los valores de Ext_{3i} son exactos en las cifras decimales incluidas, con un error real máximo de 3.68×10^{-10}. Los valores de $w_i(h = 0.1)$ se omiten en la tabla, porque ya se dieron antes.

Tabla 11.5

x_i	w_i	$y(x_i)$	$\lvert w_i - y(x_i)\rvert$
1.0	17.000000	17.000000	
1.1	15.754503	15.755455	9.520×10^{-4}
1.2	14.771740	14.773333	1.594×10^{-3}
1.3	13.995677	13.997692	2.015×10^{-3}
1.4	13.386297	13.388571	2.275×10^{-3}
1.5	12.914252	12.916667	2.414×10^{-3}
1.6	12.557538	12.560000	2.462×10^{-3}
1.7	12.299326	12.301765	2.438×10^{-3}
1.8	12.126529	12.128889	2.360×10^{-3}
1.9	12.028814	12.031053	2.239×10^{-3}
2.0	11.997915	12.000000	2.085×10^{-3}
2.1	12.027142	12.029048	1.905×10^{-3}
2.2	12.111020	12.112727	1.707×10^{-3}
2.3	12.245025	12.246522	1.497×10^{-3}
2.4	12.425388	12.426667	1.278×10^{-3}
2.5	12.648944	12.650000	1.056×10^{-3}
2.6	12.913013	12.913846	8.335×10^{-4}
2.7	13.215312	13.215926	6.142×10^{-4}
2.8	13.553885	13.554286	4.006×10^{-4}
2.9	13.927046	13.927241	1.953×10^{-4}
3.0	14.333333	14.333333	

Tabla 11.6

x_i	$w_i(h = 0.05)$	$w_i(h = 0.025)$	Ext_{1i}	Ext_{2i}	Ext_{3i}
1.0	17.00000000	17.00000000	17.00000000	17.00000000	17.00000000
1.1	15.75521721	15.75539525	15.75545543	15.75545460	15.75545455
1.2	14.77293601	14.77323407	14.77333479	14.77333342	14.77333333
1.3	13.99718996	13.99756690	13.99769413	13.99769242	13.99769231
1.4	13.38800424	13.38842973	13.38857346	13.38857156	13.38857143
1.5	12.91606471	12.91651628	12.91666881	12.91666680	12.91666667
1.6	12.55938618	12.55984665	12.56000217	12.56000014	12.56000000
1.7	12.30115670	12.30161280	12.30176684	12.30176484	12.30176471
1.8	12.12830042	12.12874287	12.12899094	12.12888902	12.12888889
1.9	12.03049438	12.03091316	12.03105457	12.03105275	12.03105263
2.0	11.99948020	11.99987013	12.00000179	12.00000011	12.00000000
2.1	12.02857252	12.02892892	12.02902924	12.02904772	12.02904762
2.2	12.11230149	12.11262089	12.11272872	12.11272736	12.11272727
2.3	12.24614846	12.24642848	12.24652299	12.24652182	12.24652174
2.4	12.42634789	12.42658702	12.42666773	12.42666673	12.42666667
2.5	12.64973666	12.64993420	12.65000086	12.65000005	12.65000000
2.6	12.91362828	12.91379422	12.91384683	12.91384620	12.91384615
2.7	13.21577275	13.21588765	13.21592641	13.21592596	13.21592593
2.8	13.55418579	13.55426075	13.55428603	13.55428573	13.55428571
2.9	13.92719268	13.92722921	13.92724153	13.92724139	13.92724138
3.0	14.33333333	14.33333333	14.33333333	14.33333333	14.33333333

CONJUNTO DE EJERCICIOS 11.4

1. Use el algoritmo no lineal de diferencias finitas con $h = 0.5$ para aproximar la solución al problema con valor de frontera

$$y'' = -(y')^2 - y + \ln x, \quad 1 \le x \le 2, \quad y(1) = 0, \quad y(2) = \ln 2.$$

Compare los resultados con la solución real $y(x) = \ln x$.

2. Use el algoritmo no lineal de diferencias finitas con $h = 0.25$ para aproximar la solución al problema con valor de frontera

$$y'' = 2y^3, \quad -1 \le x \le 0, \quad y(-1) = \frac{1}{2}, \quad y(0) = \frac{1}{3}$$

Compare los resultados con la solución real $y(x) = 1/(x + 3)$.

3. Use el algoritmo no lineal de diferencias finitas con $TOL = 10^{-4}$ para aproximar la solución a los siguientes problemas de valor de frontera. Se incluye la solución real para facilitarle la comparación de los resultados que obtenga.

a. $y'' = y^3 - yy'$, $1 \le x \le 2$, $y(1) = \frac{1}{2}$, $y(2) = \frac{1}{3}$; use $h = 0.1$ y compare el resultado para $y(x) = (x + 1)^{-1}$.

b. $y'' = 2y^3 - 6y - 2x^3$, $1 \le x \le 2$, $y(1) = 2$, $y(2) = \frac{5}{2}$; use $h = 0.1$ y compare el resultado para $y(x) = x + x^{-1}$.

c. $y'' = y' + 2(y - \ln x)^3 - x^{-1}$, $2 \le x \le 3$, $y(2) = \frac{1}{2} + \ln 2$, $y(3) = \frac{1}{3} + \ln 3$; use $h = 0.1$ y compare el resultado para $y(x) = x^{-1} + \ln x$.

d. $y'' = (x^2(y')^2 - 9y^2 + 4x^6)/x^5$, $1 \le x \le 2$, $y(1) = 0$, $y(2) = \ln 256$; use $h = 0.05$ y compare el resultado para $y(x) = x^3 \ln x$.

4. Repita el ejercicio 3(a) y (b) utilizando la extrapolación.

5. Demuestre que las hipótesis mencionadas al inicio de esta sección garantizan la no singularidad de la matriz jacobiana J para $h < 2/L$.

6. En el ejercicio 7 de la sección 11.3, aproximamos la deflexión de la viga con extremos soportados sujetos a una carga uniforme. Al usar una representación más adecuada de la curvatura se obtiene la ecuación diferencial

$$[1 + (w'(x))^2]^{-3/2} w''(x) = \frac{S}{EI} w(x) + \frac{qx}{2EI}(x - l), \quad \text{para } 0 < x < l.$$

Aproxime la deflexión $w(x)$ de la viga cada 6 plg, y después compare los resultados con los del ejercicio 7 de la sección 11.3.

11.5 El método de Rayleigh-Ritz

El método del disparo para aproximar la solución de un problema de valor de frontera sustituyó dicho problema con un par de problemas de valor inicial. El enfoque de diferencia finita reemplaza la operación continua de diferenciación con la operación discreta de diferencias finitas. El método Rayleigh-Ritz es una variante que aborda el problema desde una tercera perspectiva. Primero se reformula el problema con valor de frontera como un pro-

blema que consiste en seleccionar, del conjunto de todas las funciones suficientemente derivables que satisfacen las condiciones de frontera, aquella que reduzca al mínimo determinada integral. Después, el tamaño del conjunto de funciones factibles se disminuye, obteniéndose así una aproximación a la solución al problema de minimización y, en consecuencia, una aproximación a la solución del problema con valor de frontera.

Para describir el método de Rayleigh-Ritz consideramos la aproximación de una solución al problema lineal con valor en frontera de dos puntos, a partir del análisis del esfuerzo de una viga. Este problema con valor en frontera se describe mediante la ecuación diferencial

$$-\frac{d}{dx}\left(p(x)\,\frac{dy}{dx}\right) + q(x)y = f(x), \quad \text{para } 0 \leq x \leq 1, \tag{11.21}$$

con las condiciones de frontera

$$y(0) = y(1) = 0. \tag{11.22}$$

Esta ecuación diferencial describe la deflexión $y(x)$ de la viga de longitud 1, que tiene una sección transversal variable que está representada por $q(x)$. La deflexión se debe a los esfuerzos agregados $p(x)$ y $f(x)$.

En el análisis que sigue, supondremos que $p \in C^1[0, 1]$ y que $q, f \in C[0, 1]$. Más aún, supondremos que existe una constante $\delta > 0$ tal que

$$p(x) \geq \delta, \quad \text{tal que } q(x) \geq 0, \quad \text{para cada } x \text{ en } [0, 1].$$

Las suposiciones anteriores son suficientes para garantizar que el problema con valor en frontera de (11.22) y (11.23) tiene una solución única (véase [BSW]).

Como en el caso de los problemas con valor en frontera que describen fenómenos físicos, la solución a la ecuación de la viga satisface la propiedad **variacional**. En el caso de la ecuación de la viga, el principio variacional resulta indispensable para desarrollar el método de Rayleigh-Ritz y caracteriza la solución de esa ecuación como la función que reduce al mínimo cierta integral sobre las funciones en $C_0^2[0, 1]$, el conjunto de esas funciones u en $C^2[0, 1]$ con la propiedad de que $u(0) = u(1) = 0$. El siguiente teorema establece la caracterización.

Teorema 11.4 Sea $p \in C^1[0, 1]$, $q, f \in C[0, 1]$, y

$$p(x) \geq \delta > 0, \quad q(x) \geq 0, \quad \text{para } 0 \leq x \leq 1.$$

La función $y \in C_0^2[0, 1]$ es la solución única de la ecuación diferencial

$$-\frac{d}{dx}\left(p(x)\,\frac{dy}{dx}\right) + q(x)y = f(x), \quad \text{para } 0 \leq x \leq 1, \tag{11.23}$$

si y sólo si y es la función única en $C_0^2[0, 1]$ que reduce al mínimo la integral

$$I[u] = \int_0^1 \{p(x)[u'(x)]^2 + q(x)[u(x)]^2 - 2f(x)u(x)\}\, dx. \tag{11.24}$$

■

Los detalles de esta demostración aparecen en [Shul, páginas 88-89]. La demostración consta de tres pasos. Primero se demuestra que cualquier solución y de (11.24) satisface tambien la ecuación

$$\int_0^1 f(x)u(x)dx = \int_0^1 p(x)\frac{dy}{dx}(x)\frac{du}{dx}(x) + q(x)y(x)u(x)dx, \tag{11.25}$$

para cada $u \in C_0^2[0, 1]$.

En el segundo paso se muestra que $y \in C_0^2[0, 1]$ es una solución de (11.25) si y sólo si (11.26) se cumple para cada $u \in C_0^2[0, 1]$.

En el último paso se demuestra que (11.26) tiene una única solución, la que además será solución de (11.25) y de (11.24), de modo que las soluciones de (11.24) y (11.25) son idénticas.

Reduciendo al mínimo la integral, el método de Rayleigh-Ritz aproxima la solución y, no sobre todas las funciones en $C_0^2[0, 1]$, sino sobre un conjunto más pequeño de las que contienen combinaciones lineales de ciertas funciones básicas $\phi_1, \phi_2, \ldots, \phi_n$. Las funciones básicas son linealmente independientes y satisfacen

$$\phi_i(0) = \phi_i(1) = 0, \quad \text{para cada } i = 1, 2, \ldots, n.$$

Después de encontrar las constantes $c_1, c_2, \ldots, c_n$ que reducen al mínimo $I[\sum_{i=1}^n c_i\phi_i]$, se obtiene una aproximación $\phi(x) = \sum_{i=1}^n c_i\phi_i(x)$ a la solución $y(x)$ de la ecuación (11.24).

Conforme a la ecuación (11.25),

$$I[\phi] = I\left[\sum_{i=1}^n c_i\phi_i\right] \tag{11.26}$$

$$= \int_0^1 \left\{p(x)\left[\sum_{i=1}^n c_i\phi_i'(x)\right]^2 + q(x)\left[\sum_{i=1}^n c_i\phi_i(x)\right]^2 - 2f(x)\sum_{i=1}^n c_i\phi_i(x)\right\} dx,$$

y cuando se considera I como una función de $c_1, c_2, \ldots, c_n$ para que ocurra un mínimo es necesario tener

$$\frac{\partial I}{\partial c_j} = 0, \quad \text{para cada } j = 1, 2, \ldots, n. \tag{11.27}$$

Al derivar (11.26) se obtiene

$$\frac{\partial I}{\partial c_j} = \int_0^1 \left\{2p(x)\sum_{i=1}^n c_i\phi_i'(x)\,\phi_j'(x) + 2q(x)\sum_{i=1}^n c_i\phi_i(x)\phi_j(x) - 2f(x)\phi_j(x)\right\} dx,$$

y al sustituir en la ecuación (11.27) se obtiene

$$0 = \sum_{i=1}^n \left[\int_0^1 \{p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\}\, dx\right] c_i - \int_0^1 f(x)\phi_j(x)\, dx, \tag{11.28}$$

para toda $j = 1, 2, \ldots, n$.

Las ecuaciones descritas en la ecuación 11.28 dan como resultado un sistema lineal de $n \times n$, $A\mathbf{c} = \mathbf{b}$ en las variables $c_1, c_2, \ldots, c_n$, donde la matriz simétrica A está dada por

$$a_{ij} = \int_0^1 [p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)]\, dx,$$

y **b** se define por medio de

$$b_i = \int_0^1 f(x)\phi_i(x)\,dx.$$

La elección más elemental de las funciones básicas requiere la intervención de polinomios lineales seccionados. El primer paso consiste en formar una partición de $[0, 1]$ al escoger los puntos $x_0, x_1, \ldots, x_{n+1}$ con

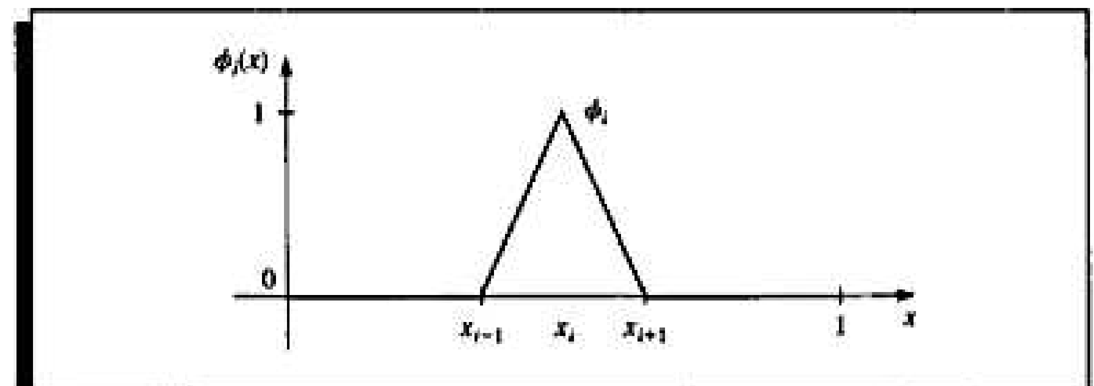
$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1.$$

Al utilizar $h_i = x_{i+1} - x_i$, para toda $i = 0, 1, \ldots, n$, definimos las funciones básicas $\phi_1(x)$, $\phi_2(x), \ldots, \phi_n(x)$ mediante

$$\phi_i(x) = \begin{cases} 0, & \text{si } 0 \le x \le x_{i-1}, \\ \dfrac{1}{h_{i-1}}(x - x_{i-1}) & \text{si } x_{i-1} < x \le x_i, \\ \dfrac{1}{h_i}(x_{i+1} - x), & \text{si } x_i < x \le x_{i+1}, \\ 0, & \text{si } x_{i+1} < x \le 1, \end{cases} \tag{11.29}$$

para toda $i = 1, 2, \ldots, n$. (Véase la Fig. 11.4.)

Figura 11.4



Las funciones ϕ_i son lineales y seccionadas; por ello, aunque las derivadas ϕ_i' no son continuas, son constantes en el subintervalo abierto (x_j, x_{j+1}) para toda $j = 0, 1, \ldots, n$. Por tanto, tenemos

$$\phi_i'(x) = \begin{cases} 0, & \text{si } 0 < x < x_{i-1}, \\ \dfrac{1}{h_{i-1}}, & \text{si } x_{i-1} < x < x_i, \\ -\dfrac{1}{h_i}, & \text{si } x_i < x < x_{i+1}, \\ 0, & \text{si } x_{i+1} < x < 1, \end{cases} \tag{11.30}$$

para toda $i = 1, 2, \ldots, n$.

Como ϕ_i y ϕ_i' son distintos de cero sólo en (x_{i-1}, x_{i+1}),

$$\phi_i(x)\,\phi_j(x) \equiv 0 \quad \text{y} \quad \phi_i'(x)\,\phi_j'(x) \equiv 0,$$

excepto cuando j es $i-1$, i, o $i+1$. En consecuencia, el sistema lineal dado por (11.29) se reduce a un sistema lineal tridiagonal de $n \times n$. Los elementos distintos de cero de A son

$$\begin{aligned} a_{ii} &= \int_0^1 \left\{p(x)[\phi_i'(x)]^2 + q(x)[\phi_i(x)]^2\right\} dx \\ &= \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\,dx + \left(\frac{-1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} p(x)\,dx \\ &\quad + \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2 q(x)\,dx + \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2 q(x)\,dx, \end{aligned}$$

para toda $i = 1, 2, \ldots, n$;

$$\begin{aligned} a_{i,i+1} &= \int_0^1 \{p(x)\phi_i'(x)\phi_{i+1}'(x) + q(x)\phi_i(x)\phi_{i+1}(x)\}\,dx \\ &= -\left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} p(x)\,dx + \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i)q(x)\,dx, \end{aligned}$$

para toda $i = 1, 2, \ldots, n-1$; y

$$\begin{aligned} a_{i,i-1} &= \int_0^1 \{p(x)\phi_i'(x)\phi_{i-1}'(x) + q(x)\phi_i(x)\phi_{i-1}(x)\}\,dx \\ &= -\left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\,dx + \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x_i - x)(x - x_{i-1})q(x)\,dx, \end{aligned}$$

para cada $i = 1, 2, \ldots, n$. Las entradas en $\mathbf{b}$ son

$$b_i = \int_0^1 f(x)\phi_i(x)dx = \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} (x - x_{i-1})f(x)\,dx + \frac{1}{h_i} \int_{x_i}^{x_{i+1}} (x_{i+1} - x)f(x)\,dx,$$

para toda $i = 1, 2, \ldots, n$.

Hay seis tipos de integrales a evaluar:

$$Q_{1,i} = \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)\,(x - x_i)q(x)\,dx, \quad \text{para cada } i = 1, 2, \ldots, n-1,$$

$$Q_{2,i} = \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2\, q(x)\,dx, \quad \text{para cada } i = 1, 2, \ldots, n,$$

$$Q_{3,i} = \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2\, q(x)\,dx, \quad \text{para cada } i = 1, 2, \ldots, n,$$

$$Q_{4,i} = \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\,dx, \quad \text{para cada } i = 1, 2, \ldots, n+1,$$

$$Q_{5,i} = \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} (x - x_{i-1}) f(x)\,dx, \quad \text{para cada } i = 1, 2, \ldots, n,$$

y

$$Q_{6,i} = \frac{1}{h_i} \int_{x_i}^{x_{i+1}} (x_{i+1} - x) f(x)\,dx, \quad \text{para cada } i = 1, 2, \ldots, n.$$

La matriz A y el vector $\mathbf{b}$ del sistema lineal $A\mathbf{c} = \mathbf{b}$ contienen los elementos

$$a_{i,i} = Q_{4,i} + Q_{4,i+1} + Q_{2,i} + Q_{3,i}, \quad \text{para cada } i = 1, 2, \ldots, n,$$
$$a_{i,i+1} = -Q_{4,i+1} + Q_{1,i}, \quad \text{para cada } i = 1, 2, \ldots, n-1,$$
$$a_{i,i-1} = -Q_{4,i} + Q_{1,i-1}, \quad \text{para cada } i = 2, 3, \ldots, n,$$

y

$$b_i = Q_{5,i} + Q_{6,i}, \quad \text{para cada } i = 1, 2, \ldots, n.$$

Los elementos de $\mathbf{c}$ son los coeficientes desconocidos $c_1, c_2, \ldots, c_n$, a partir de los cuales se construye la aproximación de Rayleigh-Ritz ϕ, dada por $\phi(x) = \sum_{i=1}^{n} c_i\phi_i(x)$.

Una dificultad práctica de este método es la necesidad de evaluar las $6n$ integrales. Pueden evaluarse directamente o mediante una fórmula de cuadratura, como el método de Simpson. Un método alternativo para la evaluación de la integral consiste en aproximar las funciones p, q y f con su polinomio interpolante lineal seccionado, e integrar luego la aproximación. Supongamos, por ejemplo, la integral $Q_{1,i}$. La interpolación lineal segmentaria de q es

$$P_q(x) = \sum_{i=0}^{n+1} q(x_i)\phi_i(x),$$

donde $\phi_1, \ldots, \phi_n$ se definen en (11.30) y

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{x_1}, & \text{si } 0 \le x \le x_1 \\ 0, & \text{en otra parte} \end{cases} \quad \text{y} \quad \phi_{n+1}(x) = \begin{cases} \dfrac{x - x_n}{1 - x_n}, & \text{si } x_n \le x \le 1 \\ 0, & \text{en otra parte.} \end{cases}$$

Dado que el intervalo de integración es $[x_i, x_{i+1}]$, $P_q(x)$ se reduce a

$$P_q(x) = q(x_i)\phi_i(x) + q(x_{i+1})\phi_{i+1}(x).$$

Este es el polinomio interpolante de primer grado que estudiamos en la sección 3.1. De acuerdo con el teorema 3.3,

$$|q(x) - P_q(x)| = O(h_i^2), \quad \text{para } x_i \le x \le x_{i+1},$$

si $q \in C^2[x_i, x_{i+1}]$. Para toda $i = 1, 2, \ldots, n-1$, la aproximación a $Q_{1,i}$ se obtiene al integrar la aproximación al integrando

$$Q_{1,i} = \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) q(x)\, dx$$

$$\approx \left(\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) \left[\frac{q(x_i)(x_{i+1} - x)}{h_i} + \frac{q(x_{i+1})(x - x_i)}{h_i}\right] dx$$

$$= \frac{h_i}{12}[q(x_i) + q(x_{i+1})].$$

Más aún, si $q \in C^2[x_i, x_{i+1}]$, entonces

$$\left|Q_{1,i} - \frac{h_i}{12}[q(x_i) + q(x_{i+1})]\right| = O(h_i^3).$$

Las aproximaciones a las otras integrales se derivan de manera parecida y están dadas por

$$Q_{2,i} \approx \frac{h_{i-1}}{12}[3q(x_i) + q(x_{i-1})], \qquad Q_{3,i} \approx \frac{h_i}{12}[3q(x_i) + q(x_{i+1})],$$

$$Q_{4,i} \approx \frac{h_{i-1}}{2}[p(x_i) + p(x_{i-1})], \qquad Q_{5,i} \approx \frac{h_{i-1}}{6}[2f(x_i) + f(x_{i-1})],$$

y

$$Q_{6,i} \approx \frac{h_i}{6}[2f(x_i) + f(x_{i+1})].$$

En el algoritmo 11.5 se establece el sistema lineal tridiagonal, y se incorpora el algoritmo 6.7 de factorización de Crout para resolver el sistema. Las integrales $Q_{1,i}, \ldots, Q_{6,i}$ pueden calcularse mediante uno de los métodos antes mencionados.

ALGORITMO 11.5

Método lineal segmentario de Rayleigh-Ritz

Para aproximar la solución al problema con valor en frontera

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y = f(x), \quad 0 \le x \le 1, \quad y(0) = y(1) = 0,$$

con la función lineal segmentaria

$$\phi(x) = \sum_{i=1}^{n} c_i \phi_i(x):$$

ENTRADA entero $n \ge 1$; puntos $x_0 = 0 < x_1 < \cdots < x_n < x_{n+1} = 1$.

SALIDA coeficientes $c_1, \ldots, c_n$.

Paso 1 Para $i = 0, \ldots, n$, tome $h_i = x_{i+1} - x_i$.

Paso 2 Para $i = 1, \ldots, n$ defina la base lineal seccionada ϕ_i por

$$\phi_i(x) = \begin{cases} 0, & 0 \le x \le x_{i-1}, \\ \dfrac{x - x_{i-1}}{h_{i-1}}, & x_{i-1} < x \le x_i, \\ \dfrac{x_{i+1} - x}{h_i}, & x_i < x \le x_{i+1}, \\ 0, & x_{i+1} < x \le 1. \end{cases}$$

Paso 3 Para $i = 1, 2, \ldots, n-1$, calcule $Q_{1,i}, Q_{2,i}, Q_{3,i}, Q_{4,i}, Q_{5,i}, Q_{6,i}$;
Calcule $Q_{2,n}, Q_{3,n}, Q_{4,n}, Q_{4,n+1}, Q_{5,n}, Q_{6,n}$.

Paso 4 Para cada $i = 1, 2, \ldots, n-1$, tome $\alpha_i = Q_{4,i} + Q_{4,i+1} + Q_{2,i} + Q_{3,i}$;
$\beta_i = Q_{1,i} - Q_{4,i+1}$;
$b_i = Q_{5,i} + Q_{6,i}$.

Paso 5 Tome $\alpha_n = Q_{4,n} + Q_{4,n+1} + Q_{2,n} + Q_{3,n}$;
$b_n = Q_{5,n} + Q_{6,n}$.

Paso 6 Tome $a_1 = \alpha_1$; **(*Los pasos* 6-10 *resuelven un sistema lineal tridiagonal simétrico utilizando el algoritmo* 6.7.)**
$\zeta_1 = \beta_1/\alpha_1$;
$z_1 = b_1/a_1$.

Paso 7 Para $i = 2, \ldots, n-1$, tome $a_i = \alpha_i - \beta_{i-1}/\zeta_{i-1}$;
$\zeta_i = \beta_i/\alpha_i$;
$z_i = (b_i - \beta_{i-1} z_{i-1})/a_i$.

Paso 8 Tome $a_n = \alpha_n - \beta_{n-1}\zeta_{n-1}$;
$z_n = (b_n - \beta_{n-1} z_{n-1})/a_n$.

Paso 9 Tome $C_n = z_n$;
SALIDA (c_n).

Paso 10 Para $i = n-1, \ldots, 1$ tome, $c_i = z_i - \zeta_i c_{i+1}$;
SALIDA (c_i).

Paso 11 PARAR. (*Procedimiento* ***terminado.***) ■

En los siguientes ejemplos **se emplea el algoritmo 11.5. Debido a lo elemental de este ejemplo, en los pasos 3, 4 y 5 las integrales se obtuvieron directamente.**

EJEMPLO 1 Considere el problema **con valor en frontera**

$$-y'' + \pi^2 y = 2\pi^2 \operatorname{sen}(\pi x), \quad 0 \le x \le 1, \quad y(0) = y(1) = 0.$$

Sea $h_i = h = 0.1$, tal que $x_i = 0.1i$ para toda $i = 0, 1, \ldots, 9$. Las integrales son

$$Q_{1,i} = 100 \int_{0.1i}^{0.1i+0.1} (0.1i + 0.1 - x)(x - 0.1i)\pi^2\, dx = \frac{\pi^2}{60},$$

$$Q_{2,i} = 100 \int_{0.1i-0.1}^{0.1i} (x - 0.1i + 0.1)^2 \pi^2 \, dx = \frac{\pi^2}{30},$$

$$Q_{3,i} = 100 \int_{0.1i}^{0.1i+0.1} (0.1i + 0.1 - x)^2 \pi^2 \, dx = \frac{\pi^2}{30},$$

$$Q_{4,i} = 100 \int_{0.1i-0.1}^{0.1i} dx = 10,$$

$$Q_{5,i} = 10 \int_{0.1i-0.1}^{0.1i} (x - 0.1i + 0.1) 2\pi^2 \operatorname{sen} \pi x \, dx$$

$$= -2\pi \cos 0.1\pi i + 20[\operatorname{sen}(0.1\,\pi i) - \operatorname{sen}((0.1i - 0.1)\,\pi)],$$

y

$$Q_{6,i} = 10 \int_{0.1i}^{0.1i+0.1} (0.1i + 0.1 - x) 2\pi^2 \operatorname{sen} \pi x \, dx$$

$$= 2\pi \cos 0.1\pi i - 20[\operatorname{sen}((0.1i + 0.1)\pi) - \operatorname{sen}(0.1\pi i)].$$

El sistema lineal $A\mathbf{c} = \mathbf{b}$ tiene

$$a_{i,i} = 20 + \frac{\pi^2}{15}, \quad \text{para toda } i = 1, 2, \ldots, 9,$$

$$a_{i,i+1} = -10 + \frac{\pi^2}{60}, \quad \text{para toda } i = 1, 2, \ldots, 8,$$

$$a_{i,i-1} = -10 + \frac{\pi^2}{60}, \quad \text{para toda } i = 1, 2, 3, \ldots, 9,$$

y

$$b_i = 40 \operatorname{sen}(0.1\pi i)[1 - \cos 0.1\pi], \quad \text{para toda } i = 1, 2, \ldots, 9.$$

La solución al sistema lineal tridiagonal es

$$c_9 = 0.3102866742,\ c_8 = 0.5902003271,\ c_7 = 0.8123410598,$$
$$c_6 = 0.9549641893,\ c_5 = 1.004108771,\ c_4 = 0.9549641893,$$
$$c_3 = 0.8123410598,\ c_2 = 0.5902003271,\ c_1 = 0.3102866742.$$

La aproximación lineal seccionada es

$$\phi(x) = \sum_{i=1}^{9} c_i \phi_i(x).$$

La solución real al problema con valor en frontera es

$$y(x) = \operatorname{sen} \pi x.$$

En la tabla 11.7 se incluye el error de la aproximación en x_i para toda $i = 1, \ldots, 9$. ■

Tabla 11.7

i	x_i	$\phi(x_i)$	$y(x_i)$	$\lvert\phi(x_i) - y(x_i)\rvert$
1	0.1	0.3102866742	0.3090169943	0.00127
2	0.2	0.5902003271	0.5877552522	0.00241
3	0.3	0.8123410598	0.8090169943	0.00332
4	0.4	0.9549641896	0.9510565162	0.00390
5	0.5	1.0041087710	1.0000000000	0.00411
6	0.6	0.9549641893	0.9510565162	0.00390
7	0.7	0.8123410598	0.8090169943	0.00332
8	0.8	0.5902003271	0.5877852522	0.00241
9	0.9	0.3102866742	0.3090169943	0.00127

Se puede demostrar que la matriz tridiagonal A dada por las funciones básicas lineales seccionada es definida positiva (véase el ejercicio 12); así que, según el teorema 6.24, el sistema lineal es estable respecto al error de redondeo. De acuerdo con las hipótesis que se mencionaron al inicio de esta sección, tenemos

$$|\phi(x) - y(x)| = O(h^2), \quad \text{para toda } x \text{ en } [0, 1].$$

Una demostración de este resultado se encuentra en [Schul, pp. 103-104].

La utilización de las funciones lineales seccionadas básicas produce una solución aproximada a las ecuaciones 11.22 y 11.23, que es continua pero no diferenciable en [0, 1]. Se requiere un conjunto más complicado de funciones básicas para construir una aproximación que pertenezca a $C_0^2[0, 1]$. Dichas funciones se parecen a los trazadores cúbicos interpolantes que se explicaron en la sección 3.4.

Recuérdese que el trazador cúbico *interpolante* S en los cinco nodos x_0, x_1, x_2, x_3 y x_4 para una función f está definido por:

a. S es un polinomio cúbico, denotado por S_j, en $[x_j, x_{j+1}]$ para toda $j = 0, 1, 2, 3$. (*Esto nos da 16 constantes seleccionables para S, 4 para cada polinomio cúbico.*)

b. $S(x_j) = f(x_j)$, para $j = 0, 1, 2, 3, 4$ (*5 condiciones especificadas*).

c. $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para $j = 0, 1, 2$ (*3 condiciones especificadas*).

d. $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$, para $j = 0, 1, 2$ (*3 condiciones especificadas*).

e. $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$, para $j = 0, 1, 2$ (*3 condiciones especificadas*).

f. Se satisface una de las siguientes condiciones de frontera:

(**i**) Libre: $S''(x_0) = S''(x_4) = 0$ (*2 condiciones especificadas*).

(**ii**) Sujeto: $S'(x_0) = f'(x_0)$ y $S'(x_4) = f'(x_4)$ (*2 condiciones especificadas*).

La unicidad de la solución requiere que el número de constantes en (a), 16, sea igual al de condiciones en (b) a (f), por lo cual solamente una de las condiciones de frontera en (f) puede especificarse para los trazadores cúbicos interpolantes.

Las funciones de trazadores cúbicos que utilizaremos en nuestras funciones básicas reciben el nombre de **trazadores B** o *trazadores en forma de campana*. Estos difieren de los trazadores interpolantes en que se satisfacen ambos conjuntos de las condiciones de frontera en (f). Para ello hay que flexibilizar dos de las condiciones de (b) a (e). Puesto que el trazador debe tener dos derivadas continuas en $[x_0, x_4]$, en la descripción de los trazadores interpolantes eliminamos dos de las condiciones de interpolación. En particular, modificamos la condición (b) y la transformamos en

b. $S(x_j) = f(x_j)$ para $j = 0, 2, 4$.

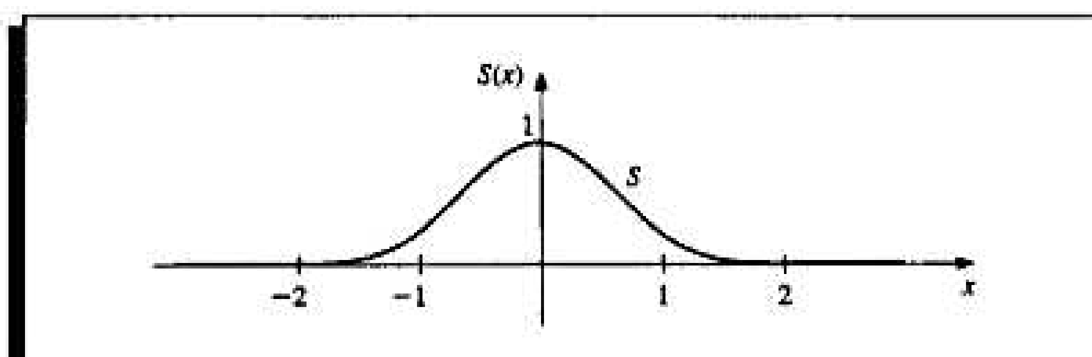
El trazador *B* básico, *S*, que se define a continuación y que aparece en la Fig. 11.5, usa los nodos uniformemente espaciados $x_0 = -2$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$ y $x_4 = 2$. Satisface las condiciones de interpolación

b. $S(x_0) = 0, \quad S(x_2) = 1, \quad S(x_4) = 0;$

y también ambos conjuntos de condiciones

$$\text{(i) } S''(x_0) = S''(x_4) = 0 \quad \text{y} \quad \text{(ii) } S'(x_0) = S'(x_4) = 0.$$

Figura 11.5



En consecuencia, $S \in C_0^2(-\infty, \infty)$, y

$$S(x) = \begin{cases} 0, & \text{si } x \leq -2, \\ \frac{1}{4}(2+x)^3, & \text{si } -2 \leq x \leq -1, \\ \frac{1}{4}\left[(2+x)^3, -4(1+x)^3\right], & \text{si } -1 < x \leq 0, \\ \frac{1}{4}\left[(2-x)^3, -4(1-x)^3\right], & \text{si } 0 < x \leq 1, \\ \frac{1}{4}(2-x)^3, & \text{si } 1 < x \leq 2, \\ 0, & \text{si } 2 < x. \end{cases} \tag{11.31}$$

Para construir las funciones básicas ϕ_i en $C_0^2[0, 1]$ primero dividimos $[0, 1]$ seleccionando un entero positivo n y definiendo $h = 1/(n+1)$. Así obtenemos los nodos uniformemente espaciados $x_i = ih$, para toda $i = 0, 1, \ldots, n+1$. Después definimos las funciones básicas $\{\phi_i\}_{i=0}^{n+1}$ como

$$\phi_i(x) = \begin{cases} S\left(\frac{x}{h}\right) - 4S\left(\frac{x+h}{h}\right), & \text{si } i = 0, \\ S\left(\frac{x-h}{h}\right) - S\left(\frac{x+h}{h}\right), & \text{si } i = 1, \\ S\left(\frac{x-ih}{h}\right), & \text{si } 2 \leq i \leq n-1, \\ S\left(\frac{x-nh}{h}\right) - S\left(\frac{x-(n+2)h}{h}\right), & \text{si } i = n, \\ S\left(\frac{x-(n+1)h}{h}\right) - 4S\left(\frac{x-(n+2)h}{h}\right), & \text{si } i = n+1. \end{cases}$$

No es difícil demostrar que $\{\phi_i\}_{i=0}^{n+1}$ es un conjunto de trazadores cúbicos linealmente independiente que satisfacen $\phi_i(0) = \phi_i(1) = 0$ para toda $i = 0, 1, \ldots, n, n+1$ (véase el ejercicio 11). Las gráficas de ϕ_i, para $2 \leq i \leq n-1$ se muestran en la Fig. 11.6 y las gráficas de ϕ_0, ϕ_1, ϕ_n y ϕ_{n+1} se pueden ver en la Fig. 11.7.

Figura 11.6

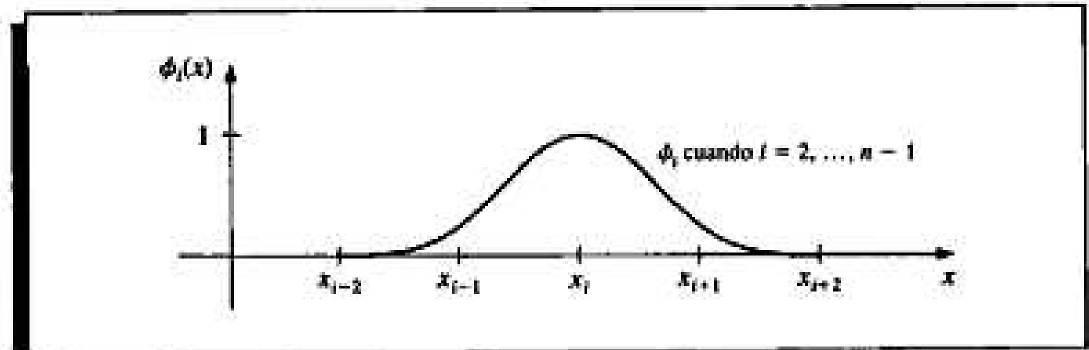
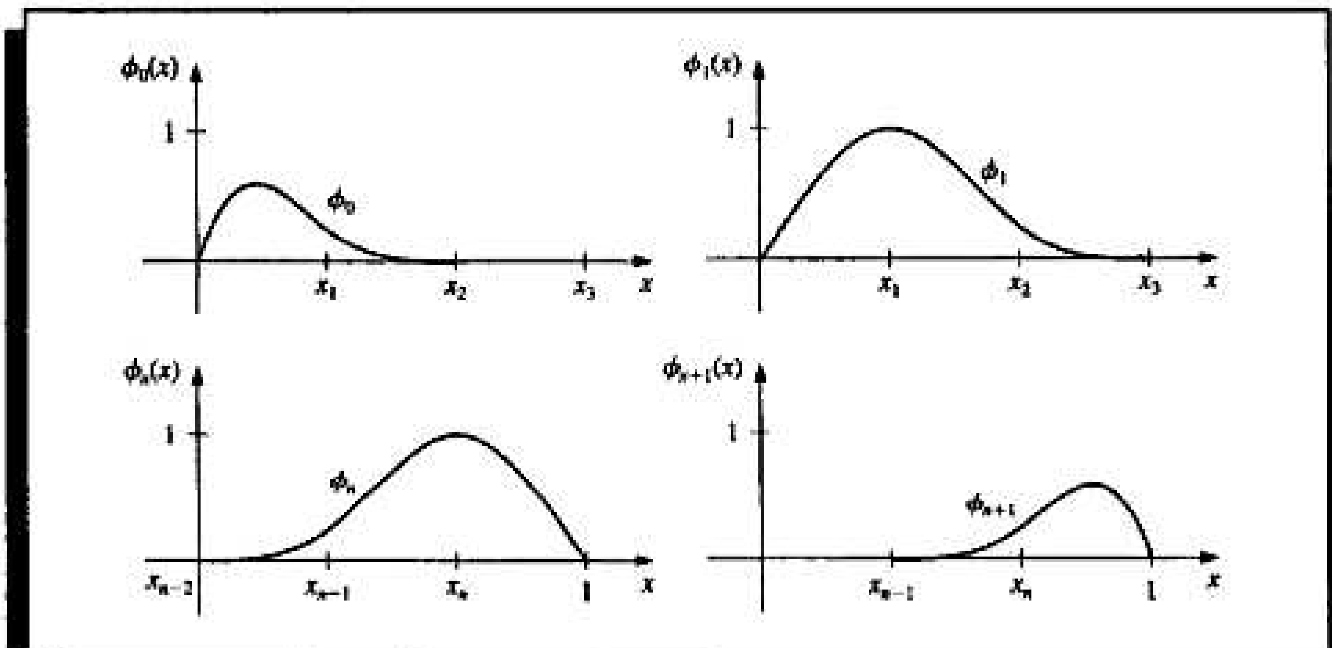


Figura 11.7



Puesto que $\phi_i(x)$ y $\phi_i'(x)$ son distintas de cero sólo para $x \in [x_{i-2}, x_{i+2}]$, la matriz de la aproximación de Rayleigh-Ritz es una matriz de banda con un ancho máximo de banda de siete:

$$A = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & 0 & \cdots & \cdots & \cdots & 0 \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & & & & \vdots \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & & & \vdots \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{n-2,n+1} \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{n-1,n+1} \\ \vdots & & & \ddots & \ddots & \ddots & \ddots & \ddots & a_{n,n+1} \\ 0 & \cdots & \cdots & \cdots & 0 & a_{n+1,n-2} & a_{n+1,n-1} & a_{n+1,n} & a_{n+1,n+1} \end{bmatrix}$$

(11.32)

donde

$$a_{ij} = \int_0^1 \{p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\,\phi_j(x)\}\,dx,$$

para toda $i, j = 0, 1, \ldots, n + 1$. El vector **b** tiene los términos

$$b_i = \int_0^1 f(x)\phi_i(x)\,dx.$$

La matriz A es definida positiva (véase el ejercicio 13) y, por tanto, podemos resolver el sistema lineal $A\mathbf{c} = \mathbf{b}$ mediante el algoritmo 6.6 de Choleski o mediante el método de la eliminación gaussiana. En el algoritmo 11.6 se describe detalladamente la construcción de la aproximación al trazador cúbico $\phi(x)$ por medio del método de Rayleigh-Ritz para el problema con valor en frontera (11.22) y (11.23) que se explicó al inicio de esa sección.

ALGORITMO 11.6

Método de trazadores cúbicos de Rayleigh-Ritz

Para aproximar la solución al problema con valor de frontera

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y = f(x), \quad 0 \le x \le 1, \quad y(0) = y(1) = 0$$

con la suma de trazadores cúbicos

$$\phi(x) = \sum_{i=0}^{n+1} c_i\phi_i(x):$$

ENTRADA entero $n \ge 1$.

SALIDA coeficientes $c_0, \ldots, c_{n+1}$.

Paso 1 Tome $h = 1/(n + 1)$.

Paso 2 Para $i = 0, \ldots, n + 1$, tome $x_i = ih$.
Tome $x_{-2} = x_{-1} = 0$; $x_{n+2} = x_{n+3} = 1$.

Paso 3 Defina la función S para

$$S(x) = \begin{cases} 0, & x \le -2, \\ \frac{1}{4}(2+x)^3, & -2 < x \le -1, \\ \frac{1}{4}\left[(2+x)^3, -4(1+x)^3\right], & -1 < x \le 0, \\ \frac{1}{4}\left[(2-x)^3, -4(1-x)^3\right], & 0 < x \le 1, \\ \frac{1}{4}(2-x)^3, & 1 < x \le 2, \\ 0, & 2 < x. \end{cases}$$

Paso 4 Defina la base del trazador cúbico $\{\phi_i\}_{i=0}^{n+1}$ por

$$\phi_0(x) = S\left(\frac{x}{h}\right) - 4S\left(\frac{x+h}{h}\right),$$

$$\phi_1(x) = S\left(\frac{x-x_1}{h}\right) - S\left(\frac{x+h}{h}\right),$$

$$\phi_i(x) = S\left(\frac{x-x_i}{h}\right), \quad \text{para } i = 2, \ldots, n-1,$$

$$\phi_n(x) = S\left(\frac{x-x_n}{h}\right) - S\left(\frac{x-(n+2)h}{h}\right),$$

$$\phi_{n+1}(x) = S\left(\frac{x-x_{n+1}}{h}\right) - 4S\left(\frac{x-(n+2)h}{h}\right).$$

Paso 5 Para $i = 0, \ldots, n+1$, haga los pasos 6-9.
(Nota: Se pueden evaluar las integrales en los pasos 6 y 9 utilizando un procedimiento de integración numérica.)

Paso 6 Para $j = i, i+1, \ldots, \text{mín}\{i+3, n+1\}$
tome $L = \text{máx}\{x_{j-2}, 0\}$;
$U = \text{mín}\{x_{i+2}, 1\}$;

$$a_{ij} = \int_L^U \left[p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\right] dx;$$

si $i \neq j$, entonces tome $a_{ji} = a_{ij}$. *(Puesto que A es simétrico.)*

Paso 7 Si $i \ge 4$, entonces para $j = 0, \ldots, i-4$ tome $a_{ij} = 0$.

Paso 8 Si $i \le n-3$, entonces para $j = i+4, \ldots, n+1$ tome $a_{ij} = 0$.

Paso 9 Tome $L = \text{máx}\{x_{i-2}, 0\}$;
$U = \text{mín}\{x_{i+2}, 1\}$;

$$b_i = \int_L^U f(x)\phi_i(x)\, dx.$$

Paso 10 Resuelva el sistema lineal $A\mathbf{c} = \mathbf{b}$, donde $A = (a_{ij})$, $\mathbf{b} = (b_0, \ldots, b_{n+1})^t$ y $\mathbf{c} = (c_0, \ldots, c_{n+1})^t$.

Paso 11 Para $i = 0, \ldots, n+1$
SALIDA (c_i).

Paso 12 PARAR. *(Procedimiento terminado.)* ■

EJEMPLO 2 Considere el problema con valor de frontera

$$-y'' + \pi^2 y = 2\,\pi^2 \operatorname{sen}(\pi x), \quad 0 \le x \le 1, \quad y(0) = y(1) = 0.$$

En el ejemplo 1 hicimos $h = 0.1$ y generamos aproximaciones usando las funciones básicas lineales seccionadas. La tabla 11.8 contiene los resultados obtenidos al aplicar los *B*-trazadores como se estableció en el algoritmo 11.6, con esta misma selección de nodos. ■

Tabla 11.8

i	c_i	x_i	$\phi(x_i)$	$y(x_i)$	$\lvert y(x_i) - \phi(x_i)\rvert$
0	$0.50964361 \times 10^{-5}$	0	0.00000000	0.00000000	0.00000000
1	0.20942608	0.1	0.30901644	0.30901699	0.00000055
2	0.39835678	0.2	0.58778549	0.58778525	0.00000024
3	0.54828946	0.3	0.80901687	0.80901699	0.00000012
4	0.64455358	0.4	0.95105667	0.95105652	0.00000015
5	0.67772340	0.5	1.00000002	1.00000000	0.00000020
6	0.64455370	0.6	0.95105713	0.95105652	0.00000061
7	0.54828951	0.7	0.80901773	0.80901699	0.00000074
8	0.39835730	0.8	0.58778690	0.58778525	0.00000165
9	0.20942593	0.9	0.30901810	0.30901699	0.00000111
10	$0.74931285 \times 10^{-5}$	1.0	0.00000000	0.00000000	0.00000000

Es recomendable que las integraciones de los pasos 6 y 9 se realicen en dos pasos. Primero, construya polinomios de trazadores cúbicos interpolantes para p, q y f mediante los métodos descritos en la sección 3.4. Después, aproxime los integrandos con productos de trazadores cúbicos o con derivadas de ellos. Ahora los integrandos son polinomios seccionados y podemos integrarlos exactamente en cada subintervalo para sumarlos después. Esto produce aproximaciones exactas de las integrales.

Las hipótesis que supusimos al inicio de esta sección son suficientes para garantizar que

$$\left\{\int_0^1 |y(x) - \phi(x)|^2\,dx\right\}^{1/2} = O(h^4), \quad \text{si } 0 \le x \le 1.$$

Una demostración de este resultado la puede encontrar en [Schul, pp. 107-108].

Los *B*-trazadores también pueden definirse para nodos espaciados de manera desigual, pero los detalles son más complicados. Una explicación de la técnica se da en [Schul, p. 73]. Otra base que se usa comúnmente la constituyen los polinomios cúbicos seccionados de Hermite. Al lector que desee una explicación excelente de este método le recomendamos consultar [Schul, pp. 24 y ss].

Otros métodos que reciben mucha atención son los de Galerkin o "de forma débil". En el caso del problema con valor en frontera que hemos venido considerando,

$$-\frac{d}{dx}\left(p(x)\,\frac{dy}{dx}\right) + q(x)y = f(x), \quad y(0) = y(1) = 0, \quad 0 \le x \le 1,$$

con las suposiciones mencionadas al inicio de esta sección, los métodos de Galerkin y de Rayleigh-Ritz se determinan por medio de la ecuación (11.29). Sin embargo, esto no se aplica a un problema de valor de frontera arbitrario. En [Schul] y [SF] se proporciona un

análisis de las semejanzas y diferencias de los dos métodos, así como una explicación de la aplicación del método de Galerkin.

Otra técnica muy común con que se resuelven los problemas con valor en frontera es el **método de colocación**. Este procedimiento comienza por seleccionar un conjunto de funciones básicas $\{\phi_1, \ldots, \phi_N\}$, un conjunto de números $\{x_i, \ldots, x_n\}$ en $[0, 1]$, y se requiere que una aproximación

$$\sum_{i=1}^{N} c_i \phi_i(x)$$

satisfaga la ecuación diferencial en los números x_j, para $1 \le j \le n$. Y si además se requiere que $\phi_i(0) = \phi_i(1) = 0$ para $1 \le i \le N$, las condiciones de frontera quedan satisfechas automáticamente. En la bibliografía especializada se ha prestado mucha atención a la elección de los números $\{x_j\}$ y a las funciones base $\{\phi_i\}$. Una elección que se presenta de manera frecuente consiste en utilizar ϕ_i como las funciones base para las funciones de trazadores relacionadas con la partición de $[0, 1]$ y tomar los nodos $\{x_j\}$ como los puntos gaussianos o las raíces de algunos polinomios ortogonales, transformados en los subintervalos apropiados. En [Ru] se incluye una comparación de varios métodos de colocación y de métodos de diferencias finitas. La conclusión es que los métodos de colocación que usan trazadores de grado superior son competitivos con los de diferencias finitas que emplean la extrapolación. Otras referencias en las que encontrará más información de este tipo de métodos son las de [DebS] y [LR].

CONJUNTO DE EJERCICIOS 11.5

1. Use el algoritmo lineal seccionado para aproximar la solución al problema con valor de frontera

$$y'' + \frac{\pi^2}{4} y = \frac{\pi^2}{16} \cos \frac{\pi}{4}x, \quad 0 \le x \le 1, \quad y(0) = y(1) = 0$$

usando $x_0 = 0$, $x_1 = 0.3$, $x_2 = 0.7$, $x_3 = 1$ y después compare los resultados con la solución real $y(x) = -\frac{1}{3} \cos \frac{\pi}{2}x - \frac{\sqrt{2}}{6} \operatorname{sen} \frac{\pi}{2}x + \frac{1}{3} \cos \frac{\pi}{4}x$.

2. Use el algoritmo lineal seccionado para aproximar la solución al problema con valor de frontera,

$$-\frac{d}{dx}(xy') + 4y = 4x^2 - 8x + 1, \quad 0 \le x \le 1, \quad y(0) = y(1) = 0$$

usando $x_0 = 0$, $x_1 = 0.4$, $x_2 = 0.8$, $x_3 = 1$ y después compare los resultados con la solución real $y(x) = x^2 - x$.

3. Use el algoritmo lineal seccionado para aproximar las soluciones a los siguientes problemas con valor de frontera y después compare los resultados con la solución real:
   a. $-x^2y'' - 2xy' + 2y = -4x^2$, $\quad 0 \le x \le 1$, $\quad y(0) = y(1) = 0$; use $h = 0.1$; solución real $y(x) = x^2 - x$.
   b. $-\frac{d}{dx}(e^x y') + e^x y = x + (2 - x)e^x$, $\quad 0 \le x \le 1$, $\quad y(0) = y(1) = 0$; use $h = 0.1$; solución real $y(x) = (x - 1)(e^{-x} - 1)$.
   c. $-\frac{d}{dx}(e^{-x} y') + e^{-x} y = (x - 1) - (x + 1)e^{-(x-1)}$, $\quad 0 \le x \le 1$, $\quad y(0) = y(1) = 0$; use $h = 0.05$; solución real $y(x) = x(e^x - e)$.
   d. $-(x + 1)y'' - y' + (x + 2)y = [2 - (x + 1)^2]e \ln 2 - 2e^x$, $\quad 0 \le x \le 1$, $\quad y(0) = y(1) = 0$; use $h = 0.05$; solución real $y(x) = e^x \ln(x + 1) - (e \ln 2)x$.

4. Use el algoritmo de trazador cúbico con $n = 3$ para aproximar la solución a los siguientes problemas con valor de frontera. Después, compare los resultados con las soluciones reales que se dan en los ejercicios 1 y 2:

   a. $y'' + \frac{\pi^2}{4}y = \frac{\pi^2}{16}\cos\frac{\pi}{4}x, \quad 0 \leq x \leq 1, \quad y(0) = 0, y(1) = 0$
   b. $-\frac{d}{dx}(xy') + 4y = 4x^2 - 8x + 1, \quad 0 \leq x \leq 1, \quad y(0) = 0, y(1) = 0$

5. Repita el ejercicio 3 usando el algoritmo del trazador cúbico.
6. Demuestre que el problema con valor de frontera

$$-\frac{d}{dx}(p(x)y') + q(x)y = f(x), \quad 0 \leq x \leq 1, \quad y(0) = \alpha, \quad y(1) = \beta,$$

   puede transformarse con un cambio de variable

$$z = y - \beta x - (1 - x)\alpha$$

   en la forma

$$-\frac{d}{dx}(p(x)z') + q(x)z = F(x), \quad 0 \leq x \leq 1, \quad z(0) = 0, z(1) = 0.$$

7. Use el ejercicio 6 y el algoritmo lineal seccionado con $n = 9$ para aproximar la solución al problema de valor de frontera

$$-y'' + y = x, \quad 0 \leq x \leq 1, \quad y(0) = 1, \quad y(1) = 1 + e^{-1}.$$

8. Repita el ejercicio 7 usando el algoritmo de trazador cúbico.
9. Demuestre que el problema con valor de frontera

$$-\frac{d}{dx}(p(x)y') + q(x)y = f(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

   puede transformarse en la forma

$$-\frac{d}{dw}(p(w)z') + q(w)z = F(w), \quad 0 \leq w \leq 1, \quad z(0) = 0, z(1) = 0,$$

   mediante un método semejante al del ejercicio 6.
10. Demuestre que las funciones base lineales a pedazos $\{\phi_i\}_{i=1}^{n}$ son linealmente independientes.
11. Demuestre que las funciones base de trazadores cúbicos $\{\phi_i\}_{i=0}^{n+1}$ son linealmente independientes.
12. Demuestre que la matriz dada por las funciones básicas lineales segmentarias es definida positiva. [*Sugerencia:* utilice la definición.]
13. Demuestre que la matriz dada por los trazadores cúbicos es definida positiva.

11.6 Reseña de métodos y de software

En este capítulo explicamos los métodos con que se aproximan las soluciones a los problemas con valor de frontera. En el caso del problema lineal con valor de frontera

$$y'' = p(x)y' + q(x)y + r(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

utilizamos el método del disparo lineal y el de diferencias finitas para aproximar la solu-

ción. En el primero se aplica el procedimiento de valor inicial para resolver los problemas

$$y'' = p(x)y' + q(x)y + r(x), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y'(a) = 0,$$

y

$$y'' = p(x)y' + q(x)y, \quad a \leq x \leq b, \quad y(a) = 0, \quad y'(a) = 1.$$

Un promedio ponderado de estas soluciones genera una solución al problema lineal con valor en frontera.

En el método de diferencias finitas, sustituimos y'' y y' con aproximaciones a la diferencia y resolvimos así un sistema lineal. Aunque las aproximaciones tal vez no sean tan exactas como las obtenidas con el método del disparo, existe menos vulnerabilidad ante el error de redondeo. Existen métodos de diferencias de orden superior, pero también podemos usar la extrapolación para mejorar la exactitud.

Para el problema de frontera no lineal

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta,$$

también presentamos dos métodos. El método de disparo no lineal requiere resolver el problema de valor inicial

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y'(a) = t,$$

en una elección inicial de t. Mejoramos la elección aplicando el método de Newton para aproximar la solución, t, a $y(b, t) = \beta$. Este método requirió resolver dos problemas de valor inicial en cada iteración. La exactitud depende de la elección del método con que se resolverán los problemas de valor inicial.

En el método de diferencias finitas para la ecuación no lineal hay que reemplazar y'' y y' por cocientes de diferencias, lo cual da origen a un sistema no lineal. Este sistema se resuelve mediante el método de Newton. Sin embargo, podemos utilizar las diferencias de orden superior o la extrapolación para mejorar la exactitud. Los métodos de diferencias finitas tienden a ser menos vulnerables al error de redondeo que los del disparo.

Ejemplificamos el método de Rayleigh-Ritz-Galerkin aproximando la solución al problema con valor en frontera

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + q(x)y = f(x), \quad 0 \leq x \leq 1, \quad y(0) = y(1) = 0.$$

Puede obtenerse una aproximación lineal seccionada o una aproximación de trazador cúbico.

La mayor parte del material referente a los problemas con valor en frontera de segundo orden puede aplicarse también a los problemas con condiciones de frontera de la forma

$$\alpha_1 y(a) + \beta_1 y'(a) = \alpha \quad \text{y} \quad \alpha_2 y(b) + \beta_2 y'(b) = \beta,$$

donde $|\alpha_1| + |\beta_1| \neq 0$ y $|\alpha_2| + |\beta_2| \neq 0$, pero algunas de las técnicas se vuelven demasiado complicadas. Al lector que le interese profundizar en este tipo de problemas le recomendamos consultar un libro que se especialice en problemas con valor en frontera, como el de [K,H].

Mencionamos sólo dos de los muchos métodos que contiene la biblioteca IMSL para resolver los problemas de valor de frontera. La subrutina BVPFD se basa en las diferencias finitas y BVPMS en los disparos múltiples, y utiliza IVPRK, método de Runge-Kutta-Verner para problemas de valor inicial. Ambas técnicas pueden emplearse con sistemas de problemas parametrizadas del valor de frontera.

La biblioteca NAG también ofrece un gran número de subrutinas para resolver los problemas con valor de frontera. La subrutina D02HAF es un método de disparo que usa el método de valor inicial de Runge-Kutta-Merson junto con el de Newton. La subrutina D02GAF usa un método de diferencias finitas con el método de Newton para resolver el sistema no lineal. La subrutina D02GBF es un método lineal de diferencias finitas y D02JAF es un método que se basa en la colocación.

Las subrutinas MUSL y MUSN, en el paquete ODE contenido en la biblioteca de Netlib, resuelven los problemas lineales y no lineales de dos puntos de frontera respectivamente. Ambos métodos están basados en métodos de disparo múltiple.

En Keller [Keller, H] y en Bailey, Shampine y Waltman[BSW] encontrará información más completa sobre los problemas generales referentes a la solución numérica de los problemas de valor de frontera de dos puntos. Roberts y Shipman [RS] centran su estudio en los métodos de disparo para este tipo de problemas, y Pryce [Pr] limita su atención a los problemas de Sturm-Liouville. El libro de Ascher, Mattheij y Russell [AMR] contiene una exposición muy amplia sobre los métodos de disparo múltiple y de disparo paralelo.

CAPÍTULO

12

Soluciones numéricas para las ecuaciones diferenciales parciales

. . .

A un cuerpo se le llama *isotrópico* si la conductividad térmica en cada uno de sus puntos es independiente de la dirección del flujo del calor a través del punto. En un cuerpo isotrópico, la temperatura, $u \equiv u(x, y, z, t)$, se obtiene resolviendo la ecuación diferencial parcial

$$\frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right) = c\rho\frac{\partial u}{\partial t},$$

donde k, c y ρ son funciones de (x, y, z) y representan, respectivamente, la conductividad térmica, el calor específico y la densidad del cuerpo en el punto (x, y, z).

Cuando k, c y ρ son constantes, a esta ecuación se le denomina ecuación simple tridimensional del calor, y se expresa como

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \frac{c\rho}{k}\frac{\partial u}{\partial t}.$$

Si la frontera del cuerpo es relativamente simple, la solución de esta ecuación se obtiene usando la serie de Fourier. En la generalidad de las situaciones donde k, c y ρ no son constantes o cuando la frontera es irregular, la solución de la ecuación diferencial parcial debe obtenerse mediante métodos de aproximación. En este capítulo ofreceremos una introducción a este tipo de técnicas.

En la sección 12.1 consideraremos la ecuación diferencial parcial **elíptica**, denominada **ecuación de Poisson**:

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y).$$

En esta ecuación suponemos que la función f describe los datos del problema en una región plana R cuya frontera denotamos con S. Este tipo de ecuaciones aparece de manera natural en el estudio de diversos problemas físicos dependientes del tiempo; por ejemplo, la distribución de calor para estado estable en una región plana, la energía potencial de un punto en un plano sobre el que operan fuerzas gravitacionales y los problemas bidimensionales del estado estable que incluyen fluidos incompresibles.

Para obtener una solución única a la ecuación de Poisson es necesario imponer otras restricciones más a la solución. Por ejemplo, el estudio de la distribución de calor para el estado estable en una región plana requiere que $f(x, y) \equiv 0$, lo cual da por resultado una simplificación de la ecuación de Poisson en

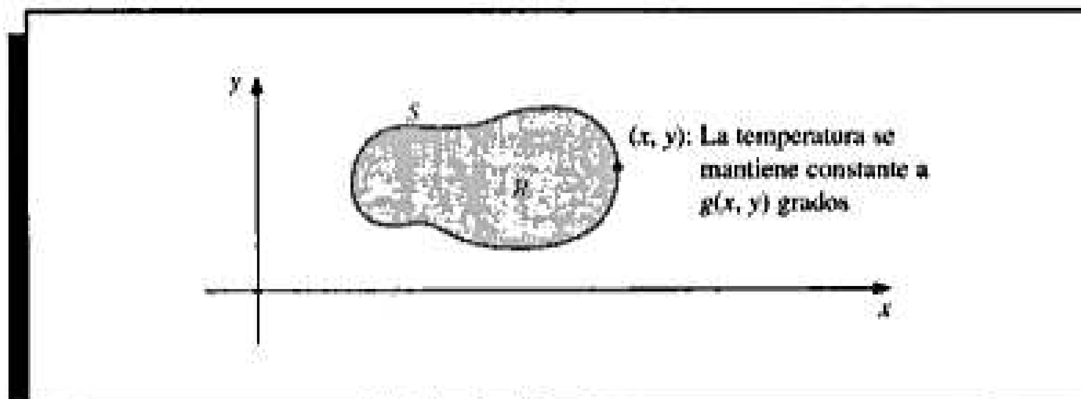
$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0,$$

que se conoce con el nombre de **ecuación de Laplace**. Si la temperatura dentro de la región está determinada por su distribución en la frontera de la región, a las restricciones se les llama **condiciones de frontera de Dirichlet**. Éstas están dadas por

$$u(x, y) = g(x, y),$$

para toda (x, y) en S, o sea, la frontera de la región R. (Véase la Fig. 12.1.)

Figura 12.1

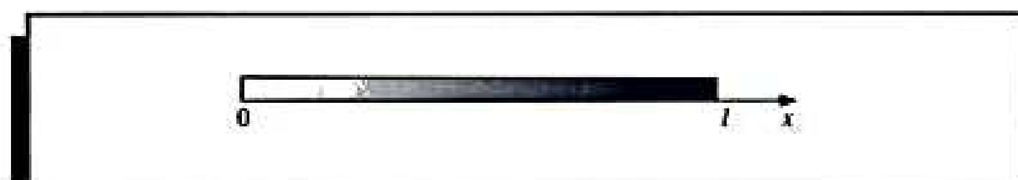


En la sección 12.2 consideraremos la solución numérica a un problema que incluye una ecuación diferencial parcial **parabólica** de la forma

$$\frac{\partial u}{\partial t}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0.$$

El problema físico considerado aquí se refiere al flujo del calor a lo largo de una barra de longitud l (véase la Fig. 12.2), la cual suponemos tiene una temperatura uniforme dentro de cada elemento transversal. Esta condición requiere que la superficie lateral de la barra esté perfectamente aislada. La constante α está determinada por las propiedades conductoras de calor del material del que está hecha la barra y se supone que es independiente de su posición en la misma.

Figura 12.2



Uno de los conjuntos comunes de restricciones en el problema del flujo del calor de este tipo consiste en especificar la distribución inicial de calor en la barra,

$$u(x, 0) = f(x),$$

y en describir el comportamiento en los extremos de la barra. Por ejemplo, si mantenemos los extremos a las temperaturas constantes U_1 y U_2, las condiciones de frontera presentarán la forma

$$u(0, t) = U_1 \quad \text{y} \quad u(l, t) = U_2,$$

y la distribución del calor en la barra se acerca a la distribución límite de la temperatura

$$\lim_{t \to \infty} u(x, t) = U_1 + \frac{U_2 - U_1}{l} x.$$

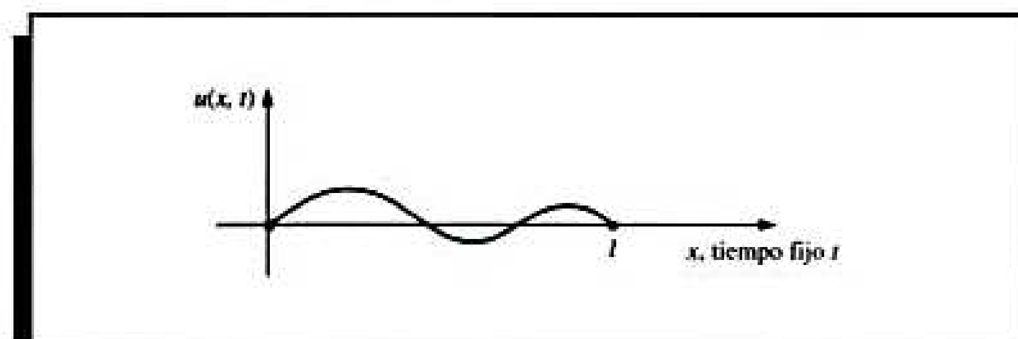
En cambio, si aislamos la barra de modo que no fluya calor por sus extremos, las condiciones de frontera serán

$$\frac{\partial u}{\partial x}(0, t) = 0 \quad \text{y} \quad \frac{\partial u}{\partial x}(l, t) = 0,$$

lo que resulta en una temperatura constante en la barra como caso límite. La ecuación diferencial parcial parabólica también es importante en el estudio de la difusión de los gases; de hecho, en algunos círculos se la conoce con el nombre de **ecuación de difusión.**

El problema que se estudió en la sección 12.3 es la **ecuación de onda** unidimensional y constituye un ejemplo de la ecuación diferencial parcial **hiperbólica**. Supóngase que alargamos una cuerda elástica de longitud l entre dos soportes al mismo nivel horizontal (véase la Fig. 12.3).

Figura 12.3



Si la ponemos en movimiento de modo que vibre en un plano vertical, el desplazamiento vertical $u(x, t)$ de un punto x en el tiempo t satisface la ecuación diferencial parcial

$$\alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = \frac{\partial^2 u}{\partial t^2}(x, t), \qquad \text{para } 0 < x < l \quad \text{y} \quad 0 < t,$$

siempre y cuando se prescinda de los efectos de amortiguamiento y la amplitud no sea demasiado grande. Para imponer restricciones a este problema, supondremos que la posición y velocidad iniciales de la cuerda están dadas por

$$u(x, 0) = f(x) \quad \text{y} \quad \frac{\partial u}{\partial t}(x, 0) = g(x), \quad \text{para } 0 \le x \le l,$$

y aplicaremos el hecho de que los extremos están fijos. Esto significa que $u(0, t) = 0$ y $u(l, t) = 0$.

Otros problemas físicos relacionados con la ecuación diferencial parcial hiperbólica se presentan en el estudio de vigas vibrantes con uno o los dos extremos sujetos, y en la transmisión de electricidad en una línea larga de transmisión donde parte de la corriente cae al suelo.

12.1 Ecuaciones diferenciales parciales elípticas

La ecuación diferencial parcial *elíptica* que estudiaremos es la ecuación de Poisson,

$$\nabla^2 u(x, y) \equiv \frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y) \tag{12.1}$$

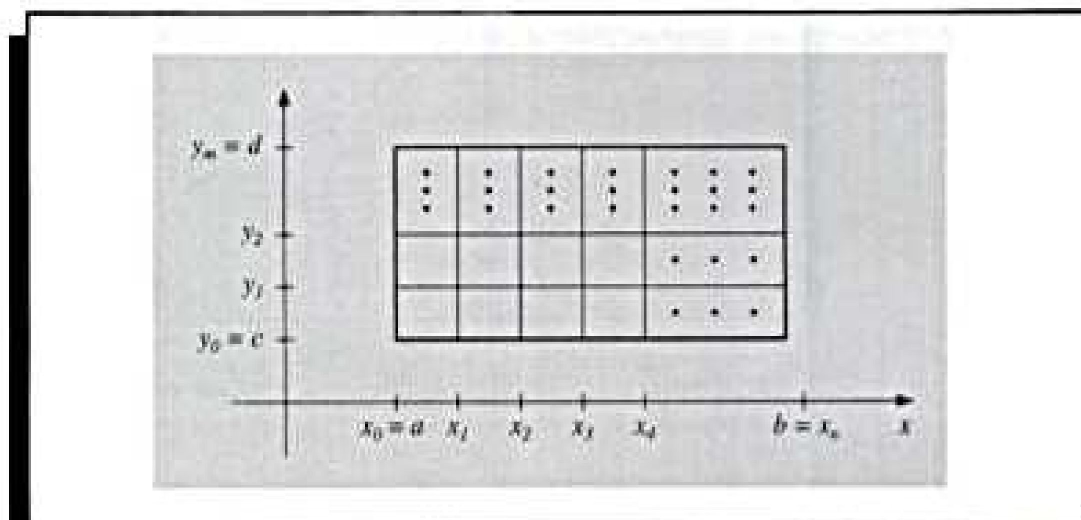
en $R = \{(x, y) \mid a < x < b, c < y < d\}$, con

$$u(x, y) = g(x, y) \quad \text{para } (x, y) \in S,$$

donde S denota la frontera de R. Para este análisis, suponemos que tanto f como g son continuas en sus dominios y que se garantiza una solución única.

El método usado es una adaptación de la técnica de diferencias finitas para problemas con valor en frontera, que se explicó en la sección 11.3. El primer paso consiste en seleccionar los enteros n y m, y en definir los tamaños de paso h y k mediante $h = (b - a)/n$ y $k = (d - c)/m$. La división del intervalo $[a, b]$ en n partes iguales de ancho h, y del intervalo $[c, d]$ en m partes iguales de ancho k (véase la Fig. 12.4), da como resultado una cua-

Figura 12.4



drícula en el rectángulo R al trazar líneas verticales y horizontales a través de los puntos con coordenadas (x_i, y_j), donde

$$x_i = a + ih, \quad \text{para cada } i = 0, 1, \ldots, n,$$

y

$$y_j = c + jk, \quad \text{para cada } j = 0, 1, \ldots, m.$$

Las líneas $x = x_i$ y $y = y_j$ son **líneas de cuadrícula,** y sus intersecciones son los **puntos de red** de la cuadrícula. En cada punto de red del interior de la cuadrícula (x_i, y_j) con $i = 1, 2, \ldots, n - 1$ y con $j = 1, 2, \ldots, m - 1$, utilizamos la serie de Taylor en la variable x alrededor de x_i para generar la fórmula de las diferencias centrales

$$\frac{\partial^2 u}{\partial x^2}(x_i, y_j) = \frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, y_j), \tag{12.2}$$

donde $\xi_i \in (x_{i-1}, x_{i+1})$. También usamos la serie de Taylor en la variable y alrededor de y_j para generar la fórmula de las diferencias centrales

$$\frac{\partial^2 u}{\partial y^2}(x_i, y_j) = \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2} - \frac{k^2}{12}\frac{\partial^4 u}{\partial y^4}(x_i, \eta_j), \tag{12.3}$$

donde $\eta_j \in (y_{j-1}, y_{j+1})$.

El uso de estas fórmulas en la ecuación (12.1) nos permite expresar la ecuación de Poisson en los puntos (x_i, y_j) como

$$\frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} + \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2}$$

$$= f(x_i, y_j) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, y_j) + \frac{k^2}{12}\frac{\partial^4 u}{\partial y^4}(x_i, \eta_j),$$

para toda $i = 1, 2, \ldots, n - 1$ y $j = 1, 2, \ldots, m - 1$, y las condiciones de frontera como

$$u(x_0, y_j) = g(x_0, y_j) \quad \text{y} \quad u(x_n, y_j) = g(x_n, y_j), \quad \text{para cada } j = 0, 1, \ldots, m;$$
$$u(x_i, y_0) = g(x_i, y_0) \quad \text{y} \quad u(x_i, y_m) = g(x_i, y_m), \quad \text{para cada } i = 1, 2, \ldots, n - 1.$$

En la forma de la ecuación de diferencias, esto da como resultado el **método de las diferencias centrales** con un error local de truncamiento del orden $O(h^2 + k^2)$:

$$2\left[\left(\frac{h}{k}\right)^2 + 1\right]w_{ij} - (w_{i+1,j} + w_{i-1,j}) - \left(\frac{h}{k}\right)^2(w_{i,j+1} + w_{i,j-1}) = -h^2 f(x_i, y_j), \tag{12.4}$$

para toda $i = 1, 2, \ldots, n - 1$ y $j = 1, 2, \ldots, m - 1$, y

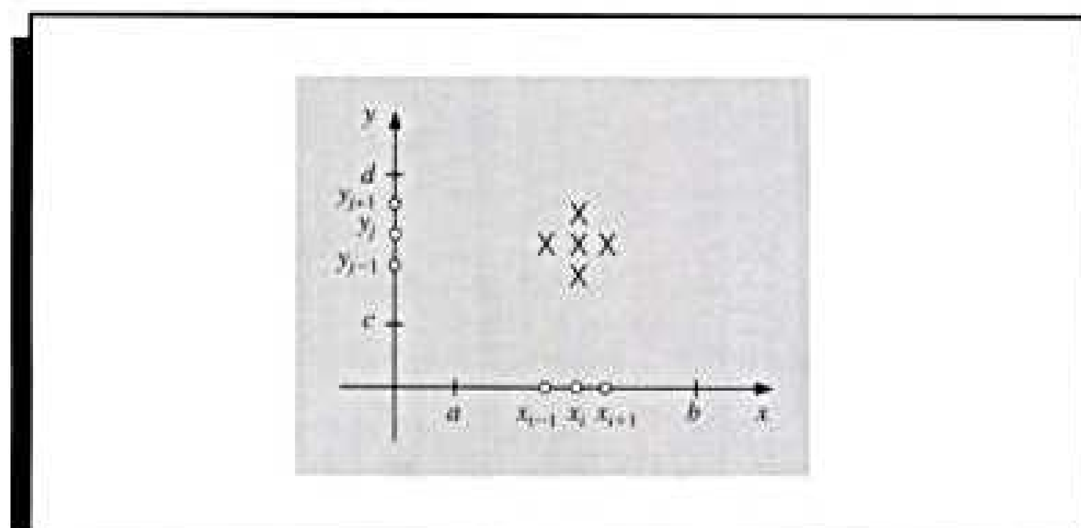
$$w_{0j} = g(x_0, y_j) \quad \text{y} \quad w_{nj} = g(x_n, y_j), \quad \text{para cada } j = 0, 1, \ldots, m; \tag{12.5}$$
$$w_{i0} = g(x_i, y_0) \quad \text{y} \quad w_{im} = g(x_i, y_m), \quad \text{para cada } i = 1, 2, \ldots, n - 1;$$

donde w_{ij} aproxima $u(x_i, y_j)$.

La ecuación común en (12.4) contiene aproximaciones a $u(x, y)$ en los puntos

$$(x_{i-1}, y_j), \quad (x_i, y_j), \quad (x_{i+1}, y_j), \quad (x_i, y_{j-1}) \quad \text{y} \quad (x_i, y_{j+1}).$$

Figura 12.5



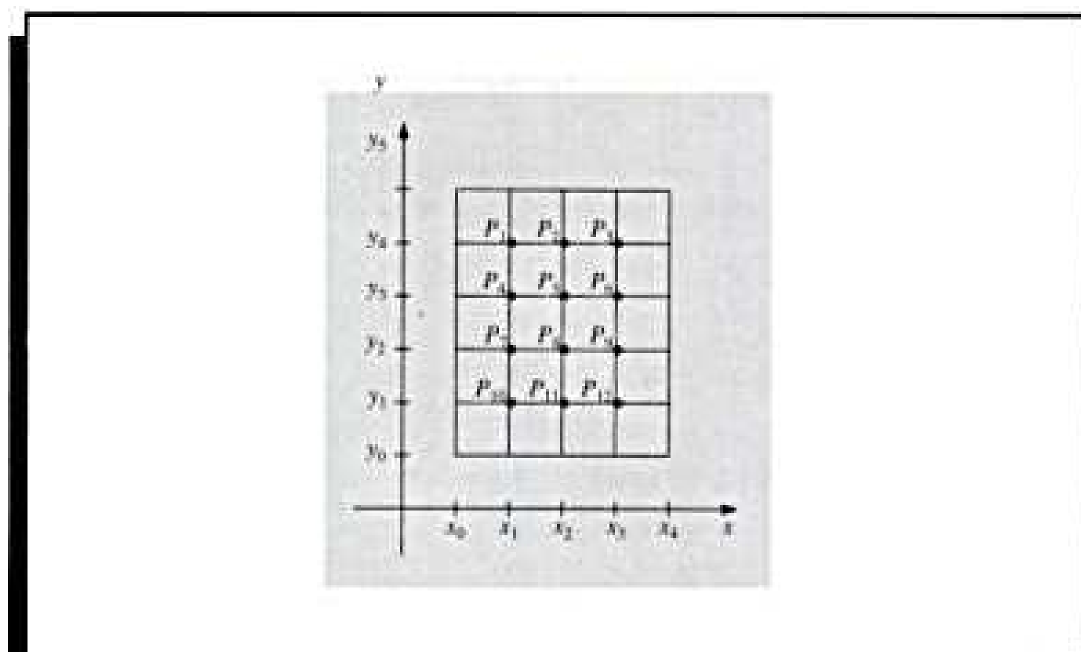
Al reproducir la parte de la cuadrícula donde estos puntos están situados (véase la Fig. 12.5), se observa que cada ecuación contiene aproximaciones en una región en forma de estrella alrededor de (x_i, y_j).

Si utilizamos la información de las condiciones de frontera (12.5) siempre que sea conveniente en el sistema dado por (12.4), es decir, en todos los puntos (x_i, y_j) adyacentes al punto de red de la frontera, tendremos un sistema lineal $(n-1)(m-1) \times (n-1)(m-1)$ cuyas incógnitas son las aproximaciones $w_{i,j}$ a $u(x_i, y_j)$ en el interior de los puntos de red.

El sistema lineal que contiene estas incógnitas se expresa más eficientemente en cálculos matriciales si se introduce un remarcaje de los puntos interiores de la red. Un sistema de marcaje de estos puntos (véase [Var, p. 187]) consiste en utilizar

$$P_l = (x_i, y_j) \quad \text{y} \quad w_l = w_{i,j},$$

Figura 12.6



donde $l = i + (m - 1 - j)(n - 1)$, para toda $i = 1, 2, \ldots, n - 1$ y $j = 1, 2, \ldots, m - 1$. Y así se marcan consecutivamente los puntos de red de izquierda a derecha y de arriba abajo. Por ejemplo, con $n = 4$ y $m = 5$, con el remarcaje se obtiene una cuadrícula cuyos puntos se muestran en la Fig. 12.6. Al marcar los puntos de este modo, se garantiza que el sistema necesario para determinar $w_{i,j}$ sea una matriz de banda con un ancho de banda máximo de $2n - 1$.

EJEMPLO 1 Considere el problema de determinar la distribución de calor en estado estable, en una placa cuadrada metálica delgada, con las dimensiones 0.5 m por 0.5 m. Conservamos dos fronteras adyacentes a 0°C, mientras el calor en las otras dos fronteras aumenta linealmente de 0°C en una esquina a 100°C en el sitio donde ambos lados se encuentran. Si ponemos los lados con las condiciones de frontera cero a lo largo de los ejes x y y, el problema se expresa así

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0,$$

para (x, y) en el conjunto $R = \{(x, y) \mid 0 < x < 0.5, 0 < y < 0.5\}$, con las condiciones de frontera

$$u(0, y) = 0, \quad u(x, 0) = 0, \quad u(x, 0.5) = 200x, \quad u(0.5, y) = 200y.$$

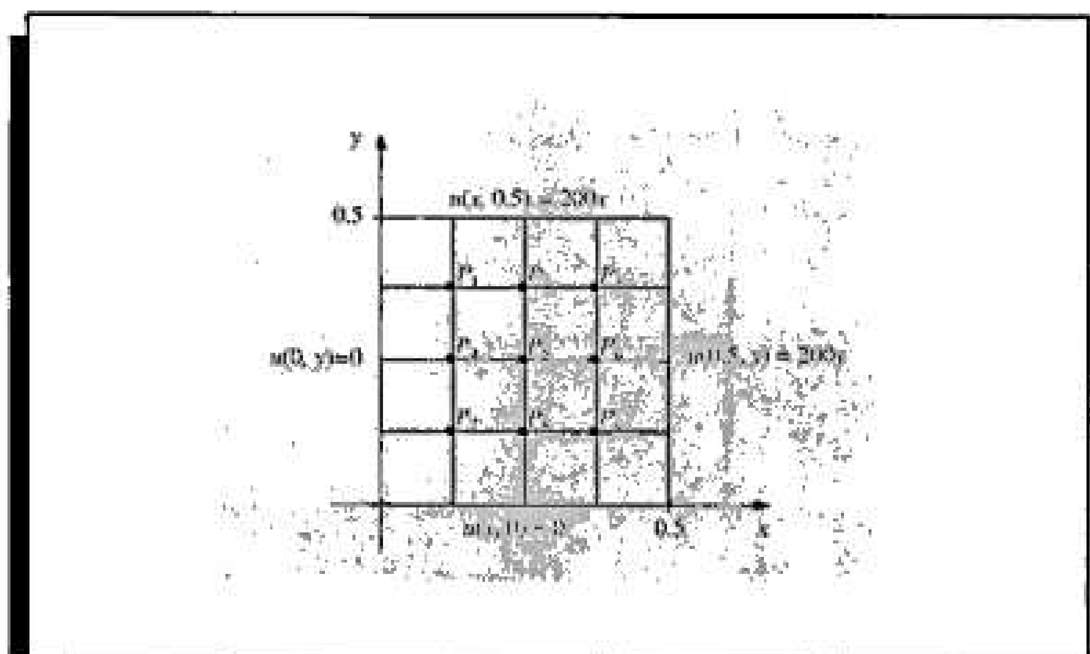
Si $n = m = 4$, el problema tiene la cuadrícula que se muestra en la Fig. 12.7, y la ecuación de diferencias (12.4) es

$$4w_{i,j} - w_{i+1,j} - w_{i-1,j} - w_{i,j-1} - w_{i,j+1} = 0,$$

para toda $i = 1, 2, 3$ y $j = 1, 2, 3$.

Expresar esto en función de los puntos remarcados de la cuadrícula interior $w_i = u(P_i)$ implica que las ecuaciones en los puntos P_i son:

Figura 12.7



$$
\begin{aligned}
P_1:&\quad 4w_1 - w_2 - w_4 = w_{0,3} + w_{1,4},\\
P_2:&\quad 4w_2 - w_3 - w_1 - w_5 = w_{2,4},\\
P_3:&\quad 4w_3 - w_2 - w_6 = w_{4,3} + w_{3,4},\\
P_4:&\quad 4w_4 - w_5 - w_1 - w_7 = w_{0,2},\\
P_5:&\quad 4w_5 - w_6 - w_4 = w_2 - w_8 = 0,\\
P_6:&\quad 4w_6 - w_5 - w_3 - w_9 = w_{4,2},\\
P_7:&\quad 4w_7 - w_8 - w_4 = w_{0,1} + w_{1,0},\\
P_8:&\quad 4w_8 - w_9 - w_7 - w_5 = w_{2,0},\\
P_9:&\quad 4w_9 - w_8 - w_6 = w_{3,0} + w_{4,1},
\end{aligned}
$$

donde los lados derechos de las ecuaciones se obtienen de las condiciones de frontera.

Las condiciones de frontera implican que

$$
\begin{aligned}
&w_{1,0} = w_{2,0} = w_{3,0} = w_{0,1} = w_{0,2} = w_{0,3} = 0,\\
&w_{1,4} = w_{4,1} = 25, \quad w_{2,4} = w_{4,2} = 50 \quad \text{y} \quad w_{3,4} = w_{4,3} = 75.
\end{aligned}
$$

El sistema lineal asociado a este problema tiene la forma

$$
\begin{bmatrix}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0\\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0\\
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0\\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0\\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1\\
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0\\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1\\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{bmatrix}
\begin{bmatrix} w_1\\ w_2\\ w_3\\ w_4\\ w_5\\ w_6\\ w_7\\ w_8\\ w_9 \end{bmatrix}
=
\begin{bmatrix} 25\\ 50\\ 150\\ 0\\ 0\\ 50\\ 0\\ 0\\ 25 \end{bmatrix}
$$

En la tabla 12.1 se muestran los valores de $w_1, w_2, \ldots, w_9$ obtenidos al aplicar a esta matriz el método de Gauss-Seidel.

Tabla 12.1

i	1	2	3	4	5	6	7	8	9
w_i	18.75	37.50	56.25	12.50	25.00	37.50	6.25	12.50	18.75

Las respuestas anteriores son exactas porque la verdadera solución, $u(x, y) = 400xy$, tiene

$$\frac{\partial^4 u}{\partial x^4} = \frac{\partial^4 u}{\partial y^4} \equiv 0,$$

y, por tanto, el error de truncamiento es cero en todos los pasos. ■

El problema que consideramos en el ejemplo 1 tiene el mismo tamaño de red, 0.125, en cada eje y requiere resolver sólo un sistema lineal de 9×9. Esto simplifica la situación, sin que origine los problemas de cálculo que surgen cuando el sistema es más grande. En

el algoritmo 12.1 se emplea el método iterativo de Gauss-Seidel para resolver el sistema lineal que se produce, y permite tamaños de red desiguales en los ejes.

ALGORITMO 12.1

Método de diferencias finitas para la ecuación de Poisson

Para aproximar la solución a la ecuación de Poisson

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y), \quad a \le x \le b, \quad c \le y \le d,$$

sujeta a las condiciones de frontera

$$u(x, y) = g(x, y) \quad \text{si } x = a \quad \text{o} \quad x = b \quad \text{y} \quad c \le y \le d$$

y

$$u(x, y) = g(x, y) \quad \text{si } y = c \quad \text{o} \quad y = d \quad \text{y} \quad a \le x \le b:$$

ENTRADA extremos a, b, c, d; enteros $m \ge 3$, $n \ge 3$; tolerancia TOL; número máximo de iteraciones N.

SALIDA aproximaciones $w_{i,j}$ a $u(x_i, y_j)$ para toda $i = 1, \ldots, n - 1$, y para toda $j = 1, \ldots, m - 1$ o un mensaje de que se excedió el número máximo de iteraciones.

Paso 1 Tome $h = (b - a)/n$;
$k = (d - c)/m$.

Paso 2 Para $i = 1, \ldots, n\ 1 - 1$, tome $x_i = a + ih$. (*En los pasos* 2 y 3 *se construyen puntos de red.*)

Paso 3 Para $j = 1, \ldots, m - 1$, tome $y_j = c + jk$.

Paso 4 Para $i = 1, \ldots, n - 1$
para $j = 1, \ldots, m - 1$, tome $w_{i,j} = 0$.

Paso 5 Tome $\lambda = h^2/k^2$;
$\mu = 2(1 + \lambda)$;
$l = 1$.

Paso 6 Mientras $l \le N$ haga los pasos 7-20. (*Los pasos 7-20 realizan iteraciones de Gauss-Seidel.*)

Paso 7 Tome $z = (-h^2 f(x_1, y_{m-1}) + g(a, y_{m-1}) + \lambda g(x_1, d) + \lambda w_{1,m-2} + w_{2,m-1})/\mu$;
$NORM = |z - w_{1,m-1}|$;
$w_{1,m-1} = z$.

Paso 8 Para $i = 2, \ldots, n - 2$
tome $z = (-h^2 f(x_i, y_{m-1}) + \lambda g(x_i, d) + w_{i-1,m-1} + w_{i+1,m-1} + \lambda w_{i,m-2})/\mu$;
si $|w_{i,m-1} - z| > NORM$, entonces tome $NORM = |w_{i,m-1} - z|$;
tome $w_{i,m-1} = z$.

Paso 9 Tome $z = (-h^2 f(x_{n-1}, y_{m-1}) + g(b, y_{m-1}) + \lambda g(x_{n-1}, d) + w_{n-2,m-1} + \lambda w_{n-1,m-2})/\mu$;

si $|w_{n-1,m-1} - z| > NORM$, entonces tome $NORM = |w_{n-1,m-1} - z|$;
tome $w_{n-1,m-1} = z$.

Paso 10 Para $j = m - 2, \ldots, 2$, haga los pasos 11, 12 y 13.

Paso 11 Tome $z = (-h^2 f(x_1, y_j) + g(a, y_j) + \lambda w_{1,j+1} + \lambda w_{1,j-1} + w_{2,j})/\mu$;
si $|w_{1,j} - z| > NORM$, entonces tome $NORM = |w_{1,j} - z|$;
tome $w_{1,j} = z$.

Paso 12 Para $i = 2, \ldots, n - 2$,
tome $z = (-h^2 f(x_i, y_j) + w_{i-1,j} + \lambda w_{i,j+1} + w_{i+1,j} + \lambda w_{i,j-1})/\mu$;
si $|w_{i,j} - z| > NORM$, entonces tome $NORM = |w_{i,j} - z|$;
tome $w_{i,j} = z$.

Paso 13 Tome $z = (-h^2 f(x_{n-1}, y_j) + g(b, y_j) + w_{n-2,j} + \lambda w_{n-1,j+1} + \lambda w_{n-1,j-1})/\mu$;
si $|w_{n-1,j} - z| > NORM$, entonces tome $NORM = |w_{n-1,j} - z|$;
tome $w_{n-1,j} = z$.

Paso 14 Tome $z = (-h^2 f(x_1, y_1) + g(a, y_1) + \lambda g(x_1, c) + \lambda w_{1,2} + w_{2,1})/\mu$;
si $|w_{1,1} - z| > NORM$, entonces tome $NORM = |w_{1,1} - z|$;
tome $w_{1,1} = z$.

Paso 15 Para $i = 2, \ldots, n - 2$,
tome $z = (-h^2 f(x_i, y_1) + \lambda g(x_i, c) + w_{i-1,1} + \lambda w_{i,2} + w_{i+1,1})/\mu$;
si $|w_{i,1} - z| > NORM$, entonces tome $NORM = |w_{i,1} - z|$;
tome $w_{i,1} = z$.

Paso 16 Tome $z = (-h^2 f(x_{n-1}, y_1) + g(b, y_1) + \lambda g(x_{n-1}, c) + w_{n-2,1} + \lambda w_{n-1,2})/\mu$;
si $|w_{n-1,1} - z| > NORM$, entonces tome $NORM = |w_{n-1,1} - z|$;
tome $w_{n-1,1} = z$.

Paso 17 Si $NORM \le TOL$, entonces haga los pasos 18 y 19.

Paso 18 Para $i = 1, \ldots, n - 1$
para $j = 1, \ldots, m - 1$ SALIDA $(x_i, y_j, w_{i,j})$.

Paso 19 PARAR. (*Procedimiento terminado con éxito.*)

Paso 20 Tome $l = l + 1$.

Paso 21 SALIDA ('Se excedió el número máximo de iteraciones');
(*Procedimiento terminado sin éxito.*)
PARAR. ■

Aunque por razones de simplicidad en el algoritmo 12.1 se incorporó el procedimiento iterativo de Gauss-Seidel, conviene utilizar una técnica directa, como la eliminación gaussiana, cuando el sistema es pequeño —del orden de 100 o menos—, pues su carácter de definida positiva garantiza la estabilidad respecto a los errores de redondeo. En particular, una generalización del algoritmo 6.7 de factorización de Crout (véase [Var, p. 221] es eficiente para resolver este sistema, ya que la matriz es simétrica en la forma tridiagonal de bloque simétrico

$$\begin{bmatrix} A_1 & C_1 & 0 & \cdots & 0 \\ C_1 & A_2 & C_2 & \ddots & \vdots \\ 0 & C_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & C_{m-1} \\ 0 & \cdots & 0 & C_{m-1} & A_{m-1} \end{bmatrix},$$

con bloques cuadrados de tamaño $(n-1) \times (n-1)$.

En el caso de sistemas grandes conviene usar un método iterativo, en concreto el método SOR, que se explica en el algoritmo 7.3. La elección de ω, que es óptima en este caso, se debe al hecho de que, cuando descomponemos A en sus partes diagonales D y en sus partes tridiagonales superior e inferior U y L,

$$A = D - L - U,$$

y B es la matriz del método de Jacobi,

$$B = D^{-1}(L + U),$$

entonces el radio espectral de B es (véase [Var, p. 203])

$$\rho(B) = \frac{1}{2}\left[\cos\left(\frac{\pi}{m}\right) + \cos\left(\frac{\pi}{n}\right)\right].$$

En consecuencia, el valor de ω que se usará es

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(B)]^2}} = \frac{4}{2 + \sqrt{4 - \left[\cos\left(\frac{\pi}{m}\right) + \cos\left(\frac{\pi}{n}\right)\right]^2}}.$$

Para lograr una convergencia más rápida del método SOR, podemos incorporar al algoritmo un método de bloques. Consúltese [Var, pp. 219-233], que explica la técnica en cuestión.

EJEMPLO 2 Consideremos la ecuación de Poisson

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = xe^y, \quad 0 < x < 2, \quad 0 < y < 1,$$

con las condiciones de frontera

$$u(0, y) = 0, \quad u(2, y) = 2e^y, \quad 0 \le y \le 1,$$
$$u(x, 0) = x, \quad u(x, 1) = ex, \quad 0 \le x \le 2.$$

Utilizaremos el algoritmo 12.1 para aproximar la solución exacta $u(x,y) = xe^y$ con $n = 6$ y con $m = 5$. El criterio de detención o paro aplicable al método de Gauss-Seidel en el paso 17 requiere que

$$\left| w_{ij}^{(l)} - w_{ij}^{(l-1)} \right| \le 10^{-10},$$

para toda $i = 1, \ldots, 5$ y $j = 1, \ldots, 4$. Por tanto, la solución de la ecuación de diferencias se obtuvo de manera exacta, y el procedimiento se detuvo en $l = 61$. El resultado, junto con los valores correctos, se incluye en la tabla 12.2. ■

Tabla 12.2

i	j	x_i	y_j	$w_{i,j}^{(61)}$	$u(x_i, y_j)$	$\lvert u(x_i, y_j) - w_{i,j}^{(61)}\rvert$
1	1	0.3333	0.2000	0.40726	0.40713	1.30×10^{-4}
1	2	0.3333	0.4000	0.49748	0.49727	2.08×10^{-4}
1	3	0.3333	0.6000	0.60760	0.60737	2.23×10^{-4}
1	4	0.3333	0.8000	0.74201	0.74185	1.60×10^{-4}
2	1	0.6667	0.2000	0.81452	0.81427	2.55×10^{-4}
2	2	0.6667	0.4000	0.99496	0.99455	4.08×10^{-4}
2	3	0.6667	0.6000	1.2152	1.2147	4.37×10^{-4}
2	4	0.6667	0.8000	1.4840	1.4837	3.15×10^{-4}
3	1	1.0000	0.2000	1.2218	1.2214	3.64×10^{-4}
3	2	1.0000	0.4000	1.4924	1.4918	5.80×10^{-4}
3	3	1.0000	0.6000	1.8227	1.8221	6.24×10^{-4}
3	4	1.0000	0.8000	2.2260	2.2255	4.51×10^{-4}
4	1	1.3333	0.2000	1.6290	1.6285	4.27×10^{-4}
4	2	1.3333	0.4000	1.9898	1.9891	6.79×10^{-4}
4	3	1.3333	0.6000	2.4302	2.4295	7.35×10^{-4}
4	4	1.3333	0.8000	2.9679	2.9674	5.40×10^{-4}
5	1	1.6667	0.2000	2.0360	2.0357	3.71×10^{-4}
5	2	1.6667	0.4000	2.4870	2.4864	5.84×10^{-4}
5	3	1.6667	0.6000	3.0375	3.0369	6.41×10^{-4}
5	4	1.6667	0.8000	3.7097	3.7092	4.89×10^{-4}

CONJUNTO DE EJERCICIOS 12.1

1. Use el algoritmo 12.1 para aproximar la solución de la ecuación diferencial parcial elíptica

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 4, \quad 0 < x < 1, \quad 0 < y < 2;$$

$$u(x, 0) = x^2, \quad u(x, 2) = (x-2)^2, \quad 0 \le x \le 1;$$

$$u(0, y) = y^2, \quad u(1, y) = (y-1)^2, \quad 0 \le y \le 2.$$

Use $h = k = \frac{1}{2}$ y compare después los resultados con la solución real $u(x, y) = (x-y)^2$.

2. Use el algoritmo 12.1 para aproximar la solución de la ecuación diferencial parcial elíptica

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad 1 < x < 2, \quad 0 < y < 1;$$

$$u(x, 0) = 2 \ln x, \quad u(x, 1) = \ln(x^2 + 1), \quad 1 \le x \le 2;$$

$$u(1, y) = \ln(y^2 + 1), \quad u(2, y) = \ln(y^2 + 4), \quad 0 \le y \le 1.$$

Use $h = k = \frac{1}{3}$ y compare después los resultados con la solución real $u(x, y) = \ln(x^2 + y^2)$.

3. Aproxime las soluciones de las siguientes ecuaciones diferenciales parciales elípticas por medio del algoritmo 12.1:

a. $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0, \quad 0 < x < 1, \quad 0 < y < 1;$

$$u(x, 0) = 0, \quad u(x, 1) = x, \quad 0 \le x \le 1;$$
$$u(0, y) = 0, \quad u(1, y) = y, \quad 0 \le y \le 1.$$

Use $h = k = 0.2$ y compare después los resultados con la solución $u(x, y) = xy$.

b. $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = -(\cos(x + y) + \cos(x - y)), \quad 0 < x < \pi, \quad 0 < y < \dfrac{\pi}{2};$

$$u(0, y) = \cos y, \quad u(\pi, y) = -\cos y, \quad 0 \le y \le \frac{\pi}{2},$$
$$u(x, 0) = \cos x, \quad u\left(x, \frac{\pi}{2}\right) = 0, \quad 0 \le x \le \pi.$$

Use $h = \pi/5$ y $k = \pi/10$ y compare después los resultados con la solución $u(x, y) = \cos x \cos y$.

c. $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = (x^2 + y^2)e^{xy}, \quad 0 < x < 2, \quad 0 < y < 1;$

$$u(0, y) = 1, \quad u(2, y) = e^{2y}, \quad 0 \le y \le 1;$$
$$u(x, 0) = 1, \quad u(x, 1) = e^{x}, \quad 0 \le x \le 2.$$

Use $h = 0.2$ y $k = 0.1$; después compare los resultados con la solución $u(x, y) = e^{xy}$.

d. $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = \dfrac{x}{y} + \dfrac{y}{x}, \quad 1 < x < 2, \quad 1 < y < 2;$

$$u(x, 1) = x \ln x, \quad u(x, 2) = x \ln(4x^2), \quad 1 \le x \le 2;$$
$$u(1, y) = y \ln y, \quad u(2, y) = 2y \ln(2y), \quad 1 \le y \le 2.$$

Use $h = k = 0.1$; y después compare los resultados con la solución $u(x, y) = xy \ln xy$.

4. Repita el ejercicio 3(a) utilizando la extrapolación con $h_0 = 0.2$, $h_1 = h_0/2$ y $h_2 = h_0/4$.

5. Construya un algoritmo semejante al 12.1, pero aplique el método SOR con ω óptima en vez del método de Gauss-Seidel para resolver el sistema lineal.

6. Repita el ejercicio 3 empleando el algoritmo construido en el ejercicio 5.

7. Un cable coaxial está hecho de un conductor interno cuadrado de 0.1 plg y un conductor externo cuadrado de 0.5 plg. La ecuación de Laplace describe el potencial en el punto de la sección transversal del cable. Suponga que conservamos el conductor externo en 0 volts y el externo en 110 volts. Calcule el potencial entre los dos conductores colocando una cuadrícula con espaciamiento horizontal de red $h = 0.1$ plg y con espaciamiento vertical de red $k = 0.1$ plg en la región

$$D = \{(x, y) \,|\, 0 \le x, y \le 0.5\}.$$

Aproxime la solución a la ecuación de Laplace en cada punto de la cuadrícula y use dos conjuntos de condiciones de frontera para derivar un sistema lineal a resolver con el método de Gauss-Seidel.

8. Una placa rectangular de plata de 6×5 cm tiene calor que se genera uniformemente en todos los puntos con una rapidez $q = 1.5$ cal/cm^3 ·s. Representemos con x la distancia a lo largo del borde de la placa de una longitud de 6 cm, y con y la distancia a lo largo del borde de la placa de una longitud de 5 cm. Supóngase que la temperatura u a lo largo de los bordes se mantiene en las siguientes temperaturas:

$$u(x, 0) = x(6 - x),\ u(x, 5) = 0, \quad 0 \le x \le 6,$$

$$u(0, y) = y(5 - y),\ u(6, y) = 0, \quad 0 \le y \le 5,$$

donde el origen se encuentra en una esquina de la placa con las coordenadas (0, 0) y los bordes se hallan a lo largo de los ejes positivos x y y. La temperatura de estado estable $u = u(x, y)$ satisface la ecuación de Poisson:

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = -\frac{q}{K}, \quad 0 < x < 6,\ 0 < y < 5,$$

donde K, la conductividad térmica, es 1.04 cal/cm·deg·s. Aproxime la temperatura $u(x, y)$ por medio del algoritmo 12.1 con $h = 0.4$ y $k = \frac{1}{3}$.

12.2 Ecuaciones diferenciales parciales parabólicas

La ecuación diferencial parcial *parabólica* que estudiaremos es la de calor o difusión

$$\frac{\partial u}{\partial t}(x, t) = \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t), \quad 0 < x < l, \quad t > 0, \tag{12.6}$$

sujeta a las condiciones

$$u(0, t) = u(l, t) = 0, \quad t > 0,$$

y

$$u(x, 0) = f(x),\ 0 \le x \le l.$$

El método que usamos para aproximar la solución de este problema contiene diferencias finitas, y se parece al que utilizamos en la sección 12.1.

Primero seleccionamos un entero $m > 0$ y sea $h = l/m$. Después seleccionamos un tamaño de peso de tiempo k. Los puntos de red para este caso son (x_i, t_j), donde $x_i = ih$ para $i = 0, 1, \ldots, m$, y $t_j = jk$, para $j = 0, 1, \ldots$

El método de diferencias se obtiene al usar la serie de Taylor en t para formar el cociente de diferencias

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_j + k) - u(x_i, t_j)}{k} - \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j), \tag{12.7}$$

para alguna $\mu_j \in (t_j, t_{j+1})$, y la serie de Taylor en x para formar el cociente de diferencias

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) = \frac{u(x_i + h, t_j) - 2u(x_i, t_j) + u(x_i - h, t_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j), \tag{12.8}$$

donde $\xi_i \in (x_{i-1}, x_{i+1})$.

La ecuación diferencial parcial parabólica (12.6) implica que en los puntos de red interiores (x_i, t_j) para toda $i = 1, 2, \ldots, m - 1$ y $j = 1, 2, \ldots,$ tendremos

$$\frac{\partial u}{\partial t}(x_i, t_j) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x_i, t_j) = 0,$$

así que el método que utilizan los cocientes de diferencias (12.7) y (12.8) es

$$\frac{w_{i,j+1} - w_{ij}}{k} - \alpha^2 \frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h^2} = 0, \tag{12.9}$$

donde w_{ij} aproxima a $u(x_i, t_j)$.

El error local de truncamiento para esta ecuación de diferencias es

$$\tau_{ij} = \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) - \alpha^2 \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j). \tag{12.10}$$

Al resolver la ecuación (12.9) para $w_{i,j+1}$ obtenemos

$$w_{i,j+1} = \left(1 - \frac{2\alpha^2 k}{h^2}\right) w_{ij} + \alpha^2 \frac{k}{h^2}(w_{i+1,j} + w_{i-1,j}), \tag{12.11}$$

para toda $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots$ Dado que la condición inicial $u(x, 0) = f(x)$, para toda $0 \leq x \leq l$, implica que $w_{i,0} = f(x_i)$, para toda $i = 0, 1, \ldots, m$, podemos usar estos valores en la ecuación (12.11) para calcular el valor de $w_{i,1}$ para toda $i = 1, 2, \ldots, m-1$. Las condiciones adicionales $u(0, t) = 0$ y $u(l, t) = 0$ implican que $w_{0,1} = w_{m,1} = 0$ y, por tanto, podemos determinar todos los elementos de la forma $w_{i,1}$. Si volvemos a aplicar el procedimiento una vez conocidas todas las aproximaciones $w_{i,1}$ podemos obtener en forma semejante los valores $w_{i,2}, w_{i,3}, \ldots$

La naturaleza explícita del método de diferencias implica que la matriz de $(m-1) \times (m-1)$ asociada a este sistema puede escribirse en la forma tridiagonal

$$A = \begin{bmatrix} (1-2\lambda) & \lambda & 0 & \cdots & 0 \\ \lambda & (1-2\lambda) & \lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \lambda \\ 0 & \cdots & 0 & \lambda & (1-2\lambda) \end{bmatrix},$$

donde $\lambda = \alpha^2(k/h^2)$. Si utilizamos

$$\mathbf{w}^{(0)} = (f(x_1), f(x_2), \ldots, f(x_{m-1}))^t$$

y

$$\mathbf{w}^{(j)} = (w_{1j}, w_{2j}, \ldots, w_{m-1,j})^t, \quad \text{para toda } j = 1, 2, \ldots,$$

entonces la solución aproximada está dada por

$$\mathbf{w}^{(j)} = A\mathbf{w}^{(j-1)}, \quad \text{para toda } j = 1, 2, \ldots$$

por tanto, $\mathbf{w}^{(j)}$ se obtiene para $\mathbf{w}^{(j-1)}$ por una matriz simple de multiplicación. A esto se le conoce con el nombre de **método de diferencias progresivas**. Si la solución a la ecuación diferencial parcial tiene cuatro derivadas parciales continuas en x y dos en t, entonces la ecuación (12.10) implica que el método es de orden $O(k + h^2)$.

EJEMPLO 1 Considere la ecuación de calor

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < 1, \quad 0 \le t,$$

con las condiciones de frontera

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$

y las condiciones iniciales

$$u(x, 0) = \operatorname{sen}(\pi x), \quad 0 \le x \le 1.$$

La solución a este problema es

$$u(x, t) = e^{-\pi^2 t} \operatorname{sen}(\pi x).$$

La solución en $t = 0.5$ será aproximada mediante el método de diferencias progresivas, primero con $h = 0.1$, $k = 0.0005$ y $\lambda = 0.05$, y después con $h = 0.1$, $k = 0.01$ y $\lambda = 1$. Los resultados se muestran en la tabla 12.3. ■

TABLA 12.3

x_i	$u(x_i, 0.5)$	$w_{i,1000}$ $k = 0.0005$	$\lvert u(x_i, 0.5) - w_{i,1000}\rvert$	$w_{i,50}$ $k = 0.01$	$\lvert u(x_i, 0.5) - w_{i,50}\rvert$
0.0	0	0	0		
0.1	0.00222241	0.00228652	6.411×10^{-5}	8.19876×10^{7}	8.199×10^{7}
0.2	0.00422728	0.00434922	1.219×10^{-4}	-1.55719×10^{8}	1.557×10^{8}
0.3	0.00581836	0.00598619	1.678×10^{-4}	2.13833×10^{8}	2.138×10^{8}
0.4	0.00683989	0.00703719	1.973×10^{-4}	-2.50642×10^{8}	2.506×10^{8}
0.5	0.00719188	0.00739934	2.075×10^{-4}	2.62685×10^{8}	2.627×10^{8}
0.6	0.00683989	0.00703719	1.973×10^{-4}	-2.49015×10^{8}	2.490×10^{8}
0.7	0.00581836	0.00598619	1.678×10^{-4}	2.11200×10^{8}	2.112×10^{8}
0.8	0.00422728	0.00434922	1.219×10^{-4}	-1.53086×10^{8}	1.531×10^{8}
0.9	0.00222241	0.00228652	6.511×10^{-5}	8.03604×10^{7}	8.036×10^{7}
1.0	0	0		0	

En el ejemplo 1 se espera un error de truncamiento del orden $O(k + h^2)$. Aunque se obtiene con $h = 0.1$ y $k = 0.0005$, esto no sucede cuando $h = 0.1$ y $k = 0.01$. Para explicar la dificultad, debemos observar la estabilidad del método de las diferencias progresivas.

Si al representar los datos iniciales se comete un error $\mathbf{e}^{(0)} = (e_1^{(0)}, e_2^{(0)}, \ldots, e_{m-1}^{(0)})^t$

$$\mathbf{w}^{(0)} = \left(f(x_1), f(x_2), \ldots, f(x_{m-1})\right)^t$$

(o si en cualquier paso la elección del paso inicial se realiza simplemente por razones de comodidad) un error de $A\mathbf{e}^{(0)}$ se propaga en $\mathbf{w}^{(1)}$, ya que

$$\mathbf{w}^{(1)} = A(\mathbf{w}^{(0)} + \mathbf{e}^{(0)}) = A\mathbf{w}^{(0)} + A\mathbf{e}^{(0)}.$$

Este proceso continúa. En el n-ésimo paso de tiempo, el error de $\mathbf{w}^{(n)}$ debido a $\mathbf{e}^{(0)}$ es $A^n\mathbf{e}^{(0)}$.

En consecuencia, el método es estable si y sólo si con cualquier error inicial $\mathbf{e}^{(0)}$ tenemos $\|A^n\mathbf{e}^{(0)}\| \le \|\mathbf{e}^{(0)}\|$ para toda n. Esto significa que $\|A^n\| \le 1$, condición que, según el teorema 7.15, requiere que el radio espectral $\rho(A^n) = (\rho(A))^n \le 1$. Así, el método de diferencias progresivas será estable sólo si $\rho(A) \le 1$.

Es posible demostrar que los valores característicos de A (véase el ejercicio 7) son

$$\mu_i = 1 - 4\lambda\left(\operatorname{sen}\left(\frac{i\pi}{2m}\right)\right)^2, \quad \text{para cada } i = 1, 2, \ldots, m-1.$$

Por tanto, la condición de estabilidad se reduce a determinar si

$$\rho(A) = \max_{1\le i\le m-1}\left|1 - 4\lambda\left(\operatorname{sen}\frac{i\pi}{2m}\right)^2\right| \le 1,$$

lo cual se simplifica y transforma en

$$0 \le \lambda\left(\operatorname{sen}\left(\frac{i\pi}{2m}\right)\right)^2 \le \frac{1}{2}, \quad \text{para cada } i = 1, 2, \ldots, m-1.$$

Se requiere que esta condición de desigualdad se conserve cuando $h \to 0$, o, en forma equivalente, cuando $m \to \infty$; por ello el hecho de que

$$\lim_{m\to\infty}\left[\operatorname{sen}\left(\frac{(m-1)\pi}{2m}\right)\right]^2 = 1$$

significa que la estabilidad ocurrirá sólo si $0 \le \lambda \le \frac{1}{2}$. Puesto que $\lambda = \alpha^2(k/h^2)$, esta desigualdad requiere elegir h y k, de modo que

$$\alpha^2\frac{k}{h^2} \le \frac{1}{2}.$$

En el ejemplo 1, $\alpha = 1$, de modo que esta condición se cumple cuando $h = 0.1$ y $k = 0.0005$; pero cuando aumentamos k a 0.01 sin un incremento de h correspondiente, la razón fue

$$\frac{0.01}{(0.1)^2} = 1 > \frac{1}{2},$$

manifestándose problemas de estabilidad.

Para ser consistentes con la terminología del capítulo 5, llamamos **condicionalmente estable** al método de diferencias progresivas, y hacemos énfasis en que converge a la solución de la ecuación (12.6) con la rapidez de convergencia $O(k + h^2)$, a condición de que

$$\alpha^2\frac{k}{h^2} \le \frac{1}{2}$$

y se cumplen las condiciones requeridas de continuidad. (Véase en [IK, pp. 502-505] una demostración detallada de este hecho.)

Para obtener un método que sea **incondicionalmente estable,** consideraremos un método de diferencias implícitas que se obtiene al usar el cociente de diferencias regresivas para $(\partial u/\partial t)(x_i, t_j)$, en la forma

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_j) - u(x_i, t_{j-1})}{k} + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j),$$

donde μ_j está en (t_{j-1}, t_j). Al sustituir esta ecuación, junto con la ecuación (12.8) para $\partial^2 u/\partial x^2$, en la ecuación diferencial parcial, obtenemos

$$\frac{u(x_i, t_j) - u(x_i, t_{j-1})}{k} - \alpha^2 \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$

$$= -\frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) - \alpha^2 \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j),$$

para alguna $\xi_i \in (x_{i-1}, x_{i+1})$. El **método de diferencias regresivas** que resulta es

$$\frac{w_{ij} - w_{i,j-1}}{k} - \alpha^2 \frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h^2} = 0, \tag{12.12}$$

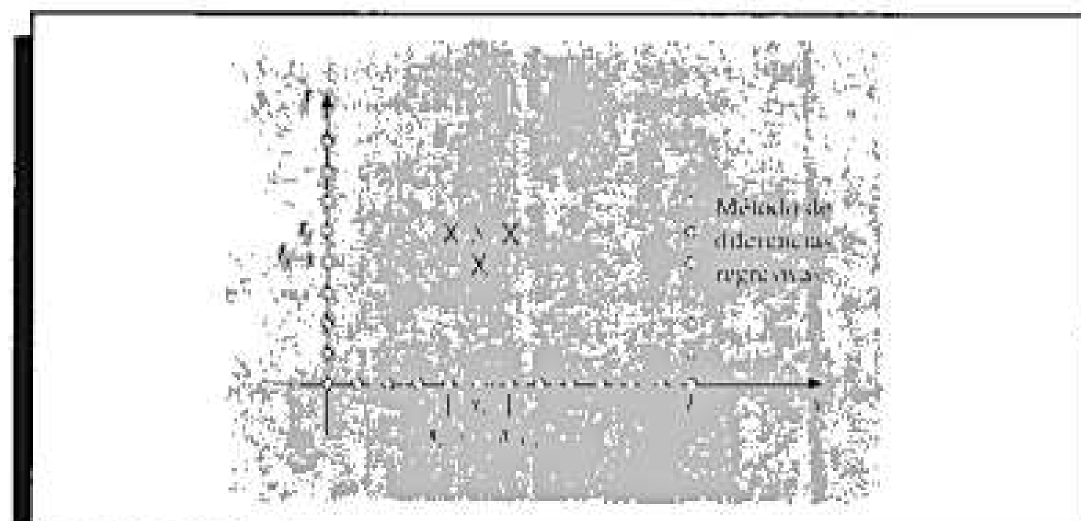
para toda $i = 1, 2, \ldots, m-1$, y $j = 1, 2, \ldots$

El método de diferencias regresivas incluye, en un paso típico, los puntos de red

$$(x_i, t_j), \quad (x_i, t_{j-1}), \quad (x_{i-1}, t_j) \quad \text{y} \quad (x_{i+1}, t_j),$$

y, en forma de rejilla o cuadrícula, contiene aproximaciones en los puntos marcados con × en la Fig. 12.8.

Figura 12.8



Las condiciones iniciales y de frontera relacionadas con el problema suministran información en los puntos circulados de red; por ello, la figura muestra que no es posible utilizar procedimientos explícitos para resolver la ecuación (12.12). En el método de diferencias progresivas (véase la Fig. 12.9) se utilizaron las aproximaciones en

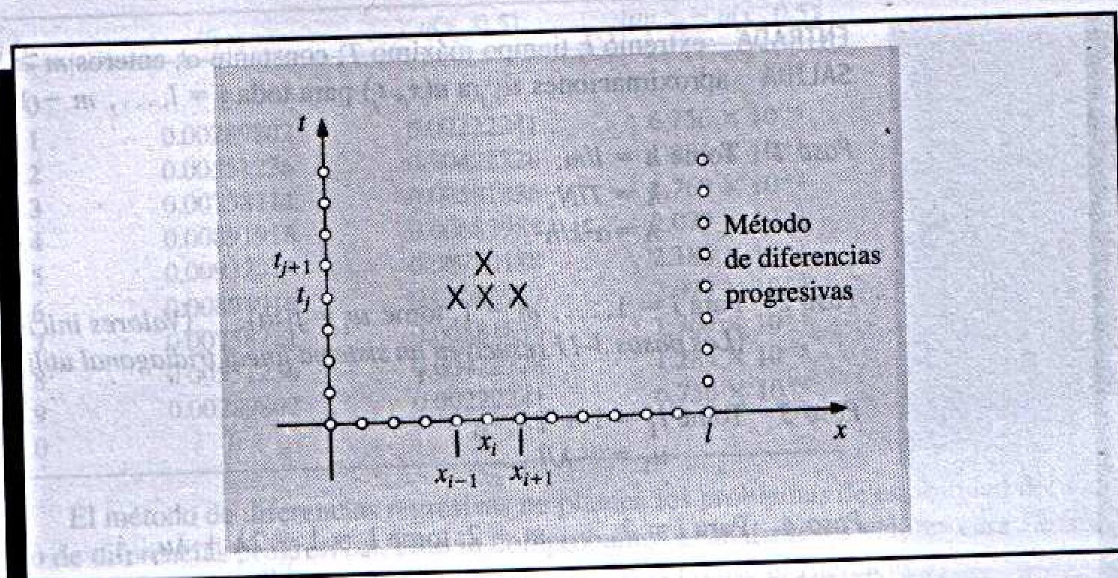
$$(x_{i-1}, t_j), \quad (x_i, t_j), \quad (x_i, t_{j+1}) \quad \text{y} \quad (x_{i+1}, t_j),$$

de modo que dispusimos de un método explícito para calcular las aproximaciones, que tenía como base la información proveniente de las condiciones iniciales y de frontera.

Si una vez más denotamos con λ la cantidad $\alpha^2(k/h^2)$, el método de diferencias regresivas se convierte en

$$(1 + 2\lambda)w_{ij} - \lambda w_{i+1,j} - \lambda w_{i-1,j} = w_{i,j-1},$$

Figura 12.9



para toda $i = 1, 2, \ldots, m-1$, y $j = 1, 2, \ldots$ Aplicando el hecho de que $w_{i,0} = f(x_i)$ para toda $i = 1, 2, \ldots, m-1$ y $w_{m,j} = w_{0,j} = 0$ para toda $j = 1, 2, \ldots$, este método de diferencias tiene la representación matricial:

$$\begin{bmatrix} (1+2\lambda) & -\lambda & 0 & \cdots & \cdots & 0 \\ -\lambda & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -\lambda \\ 0 & \cdots & \cdots & 0 & -\lambda & (1+2\lambda) \end{bmatrix} \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{m-1,j} \end{bmatrix} = \begin{bmatrix} w_{1,j-1} \\ w_{2,j-1} \\ \vdots \\ w_{m-1,j-1} \end{bmatrix}, \qquad (12.13)$$

o $A\mathbf{w}^{(j)} = \mathbf{w}^{(j-1)}$ para toda $j = 1, 2, \ldots$

Así, debemos resolver ahora un sistema lineal para obtener $\mathbf{w}^{(j)}$ a partir de $\mathbf{w}^{(j-1)}$. Dado que $\lambda > 0$, la matriz A es definida positiva y estrictamente dominante en forma diagonal, además de ser tridiagonal. Para resolver este sistema, podemos emplear la factorización de Crout para sistemas lineales tridiagonales del algoritmo 6.7, o el algoritmo SOR 7.3. El algoritmo 12.2 resuelve (12.13) mediante la factorización de Crout, que es un método aceptable a menos que m sea grande. En este algoritmo suponemos, para propósitos de detención o paro, que se da una cota para t.

ALGORITMO 12.2

Método de diferencias regresivas para la ecuación de calor

Para aproximar la solución a la ecuación diferencial parcial parabólica

$$\frac{\partial u}{\partial t}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < l, \quad 0 < t < T,$$

sujeta a las condiciones de frontera

$$u(0, t) = u(l, t) = 0, \quad 0 < t < T,$$

y a las condiciones iniciales

$$u(x, 0) = f(x), \quad 0 \le x \le l:$$

ENTRADA extremo l; tiempo máximo T; constante α; enteros $m \geq 3$, $N \geq 1$.
SALIDA aproximaciones $w_{i,j}$ a $u(x_i, t_j)$ para toda $i = 1, \ldots, m-1$ y $j = 1, \ldots, N$.

Paso 1 Tome $h = l/m$;
$k = T/N$;
$\lambda = \alpha^2 k/h^2$.

Paso 2 Para $i = 1, \ldots, m-1$, tome $w_i = f(ih)$. (*Valores iniciales.*)
(*Los pasos 3-11 resuelven un sistema lineal tridiagonal utilizando el Algoritmo 6.7.*)

Paso 3 Tome $l_1 = 1 + 2\lambda$;
$u_1 = -\lambda/l_1$.

Paso 4 Para $i = 2, \ldots, m-2$, tome $l_i = 1 + 2\lambda + \lambda u_{i-1}$;
$u_i = -\lambda/l_i$.

Paso 5 Tome $l_{m-1} = 1 + 2\lambda + \lambda u_{m-2}$.

Paso 6 Para $j = 1, \ldots, N$, haga los pasos 7-11.

Paso 7 Tome $t = jk$; (t_j *actual.*)
$z_1 = w_1/l_1$.

Paso 8 Para $i = 2, \ldots, m-1$, tome $z_i = (w_i + \lambda z_{i-1})/l_i$.

Paso 9 Tome $w_{m-1} = z_{m-1}$.

Paso 10 Para $i = m-2, \ldots, 1$, tome $w_i = z_i - u_i w_{i+1}$.

Paso 11 SALIDA (t); (*Nota:* $t = t_j$.)
Para $i = 1, \ldots, m-1$, tome $x = ih$;
SALIDA (x, w_i). (*Nota:* $w_i = w_{i,j}$.)

Paso 12 PARAR. (*Procedimiento terminado.*) ■

EJEMPLO 2 Utilizamos el método de diferencias regresivas (algoritmo 12.2) con $h = 0.1$ y con $k = 0.01$ para aproximar la solución de la ecuación de calor

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < 1, \quad 0 < t,$$

sujeta a las restricciones

$$u(0, t) = u(1, t) = 0, \quad 0 < t, \quad u(x, 0) = \operatorname{sen} \pi x, \quad 0 \leq x \leq 1,$$

que se consideró en el ejemplo 1. Para demostrar la estabilidad incondicional de este método, volvemos a comparar $w_{i,50}$ con $u(x_i, 0.5)$, donde $i = 0, 1, \ldots, 10$.

Los resultados de la tabla 12.4 tienen los mismos valores de h y k que los de las columnas quinta y sexta de la 12.3, lo que ilustra la estabilidad de este método. ■

Tabla 12.4

x_i	$w_{i,50}$	$u(x_i, 0.5)$	$\lvert w_{i,50} - u(x_i, 0.5)\rvert$
0.0	0	0	
0.1	0.00289802	0.00222241	6.756×10^{-4}
0.2	0.00551236	0.00422728	1.285×10^{-3}
0.3	0.00758711	0.00581836	1.769×10^{-3}
0.4	0.00891918	0.00683989	2.079×10^{-3}
0.5	0.00937818	0.00719188	2.186×10^{-3}
0.6	0.00891918	0.00683989	2.079×10^{-3}
0.7	0.00758711	0.00581836	1.769×10^{-3}
0.8	0.00551236	0.00422728	1.285×10^{-3}
0.9	0.00289802	0.00222241	6.756×10^{-4}
1.0	0	0	

El método de diferencias regresivas no plantea los problemas de estabilidad del método de diferencias progresivas. Esto lo comprobamos al analizar los valores característicos de la matriz. En el método de diferencias regresivas (véase el Ejercicio 8) los valores característicos son

$$\mu_i = 1 + 4\lambda \left[\operatorname{sen}\left(\frac{i\pi}{2m}\right)\right]^2, \quad \text{para cada } i = 1, 2, \ldots, m-1,$$

y como $\lambda > 0$, tendremos $\mu_i > 1$ para toda $i = 1, 2, \ldots, m-1$. Esto implica que existe A^{-1} porque cero no es un valor característico de A. Un error $\mathbf{e}^{(0)}$ en los datos iniciales genera un error $(A^{-1})^n\mathbf{e}^{(0)}$ en el n-ésimo paso. Y como los valores característicos de A^{-1} son los recíprocos de los valores característicos de A, el radio espectral de A^{-1} está acotado superiormente por 1 y el método es estable, independientemente de la elección de $\lambda = \alpha^2(k/h^2)$. En la terminología usada en el capítulo 5, llamamos **incondicionalmente estable** al método de diferencias regresivas. El error local de truncamiento de esta técnica es del orden de $O(k + h^2)$, siempre y cuando la solución de la ecuación diferencial satisfaga las condiciones normales de diferenciabilidad. En este caso, el método converge a la solución de la ecuación diferencial parcial con la misma rapidez (véase a [IK, p. 508]).

La debilidad del método de diferencias regresivas radica en el hecho de que el error local de truncamiento tiene una parte con orden $O(k)$, la cual requiere hacer mucho más pequeños los intervalos de tiempo que los de espacio. Sin duda convendría contar con un procedimiento cuyo error local de truncamiento fuese de $O(k^2 + h^2)$. El primer paso en esta dirección consiste en emplear una ecuación de diferencias que tenga un error de $O(k^2)$ para $u_t(x, t)$ en vez de las que hemos usado antes, cuyo error fue de $O(k)$. Esto podemos hacerlo utilizando la serie de Taylor en t para la función $u(x, t)$ en el punto (x_i, t_j) y evaluando después en (x_i, t_{j+1}) y en (x_i, t_{j-1}) para obtener la fórmula de las diferencias centrales

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_{j+1}) - u(x_i, t_{j-1})}{2k} + \frac{k^2}{6}\frac{\partial^3 u}{\partial t^3}(x_i, \mu_j),$$

donde $\mu_j \in (t_{j-1}, t_{j+1})$. El método de diferencias que resulta al sustituir esto y el cociente común de diferencias para $(\partial^2 u/\partial x^2)$, ecuación (12.8), en la ecuación diferencial, recibe el nombre de **método de Richardson** y está dado por

$$\frac{w_{i,j+1} - w_{i,j-1}}{2k} - \alpha^2 \frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h^2} = 0. \tag{12.14}$$

El método de Richardson tiene un error local de truncamiento del orden $O(k^2 + h^2)$, pero lamentablemente también presenta serios problemas de estabilidad (véase el ejercicio 6).

Un método más prometedor se deriva al promediar el método de diferencias progresivas en el j-ésimo paso en t,

$$\frac{w_{i,j+1} - w_{i,j}}{k} - \alpha^2 \frac{w_{i+1,j} - 2_{w_{i,j}} + w_{i-1,j}}{h^2} = 0,$$

que tiene el error local de truncamiento

$$\tau_F = \frac{k}{2} \frac{\partial^2 u}{\partial t^2} (x_i, \mu_j) + O(h^2),$$

y el método de diferencias regresivas en el $(j + 1)$-ésimo paso en t,

$$\frac{w_{i,j+1} - w_{i,j}}{k} - \alpha^2 \frac{w_{i+1,j+1} - 2w_{i,j+1} + w_{i-1,j+1}}{h^2} = 0,$$

que tiene un error local de truncamiento

$$\tau_B = -\frac{k}{2} \frac{\partial^2 u}{\partial t^2} (x_i, \hat{\mu}_j) + O(h^2).$$

Si suponemos que

$$\frac{\partial^2 u}{\partial t^2} (x_i, \hat{\mu}_j) \approx \frac{\partial^2 u}{\partial t^2} (x_i, \mu_j),$$

entonces el método de diferencias promediadas,

$$\frac{w_{i,j+1} - w_{ij}}{k} - \frac{\alpha^2}{2}\left[\frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} + \frac{w_{i+1,j+1} - 2w_{i,j+1} + w_{i-1,j+1}}{h^2}\right] = 0,$$

tiene un error local de truncamiento del orden $O(k^2 + h^2)$, siempre y cuando se cumplan las condiciones normales de diferenciabilidad. A esto se le llama **método de Crank-Nicolson** y está representado en la forma matricial

$$A\mathbf{w}^{(j+1)} = B\mathbf{w}^{(j)}, \quad \text{para cada } j = 0, 1, 2, \ldots \tag{12.15}$$

donde

$$\lambda = \alpha^2 \frac{k}{h^2}, \quad \mathbf{w}^{(j)} = (w_{1,j}, w_{2,j}, \ldots, w_{m-1,j})^t,$$

y las matrices A y B están dadas por:

$$A = \begin{bmatrix} (1+\lambda) & -\frac{\lambda}{2} & 0 & \cdots & 0 \\ -\frac{\lambda}{2} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -\frac{\lambda}{2} \\ 0 & \cdots & 0 & -\frac{\lambda}{2} & (1+\lambda) \end{bmatrix}$$

y

$$B = \begin{bmatrix} (1-\lambda) & \frac{\lambda}{2} & 0 & \cdots & 0 \\ \frac{\lambda}{2} & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{\lambda}{2} \\ 0 & \cdots & 0 & \frac{\lambda}{2} & (1-\lambda) \end{bmatrix}.$$

Por tanto, A es no singular, porque es una matriz definida positiva, estrictamente dominante en forma diagonal y tridiagonal. Podemos usar la factorización de Crout para un sistema lineal tridiagonal del algoritmo 6.7 o el algoritmo SOR 7.3 para obtener $\mathbf{w}^{(j+1)}$ a partir de $\mathbf{w}^{(j)}$ para toda $j = 0, 1, 2, \ldots$ El algoritmo 12.3 incorpora la factorización de Crout en el método de Crank-Nicolson. Al igual que en el caso del algoritmo 12.2, si queremos determinar un procedimiento de paro o detención debe especificarse una longitud finita para el intervalo de tiempo. En [IK, pp. 508-512] podrá encontrar la comprobación de que el método de Crank-Nicolson es incondicionalmente estable y tiene el orden de convergencia $O(k^2 + h^2)$.

ALGORITMO **12.3**

Método de Crank-Nicolson

Para aproximar la solución de la ecuación diferencial parcial parabólica

$$\frac{\partial u}{\partial t}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < l, \quad 0 < t < T,$$

sujeta a las condiciones de frontera

$$u(0, t) = u(l, t) = 0, \quad 0 < t < T,$$

y a las condiciones iniciales

$$u(x, 0) = f(x), \quad 0 \leq x \leq l:$$

ENTRADA extremo l; tiempo máximo T; constante α; enteros $m \geq 3$, $N \geq 1$.

SALIDA aproximaciones $w_{i,j}$ a $u(x_i, t_j)$ para toda $i = 1, \ldots, m-1$ y $j = 1, \ldots, N$.

Paso 1 Tome $h = l/m$;
$k = T/N$;
$\lambda = \alpha^2 k/h^2$.
$w_m = 0$.

Paso 2 Para $i = 1, \ldots, m-1$, tome $w_i = f(ih)$. (*Valores iniciales.*)
(*Los pasos 3-11 resuelven un sistema lineal tridiagonal utilizando el Algoritmo 6.7.*)

Paso 3 Tome $l_1 = 1 + \lambda$;
$u_1 = -\lambda/(2l_1)$.

Paso 4 Para $i = 2, \ldots, m-2$, tome $l_i = 1 + \lambda + \lambda u_{i-1}/2$;
$u_i = -\lambda/(2l_i)$.

Paso 5 Tome $l_{m-1} = 1 + \lambda + \lambda u_{m-2}/2$.

Paso 6 Para $j = 1, \ldots, N$ haga los pasos 7-11.

Paso 7 Tome $t = jk$; (t_j *actual.*)

$$z_1 = \left[(1-\lambda)w_1 + \frac{\lambda}{2}w_2\right] \Big/ l_1.$$

Paso 8 Para $i = 2, \ldots, m-1$, tome

$$z_i = \left[(1-\lambda)w_i + \frac{\lambda}{2}(w_{i+1} + w_{i-1} + z_{i-1})\right] \Big/ l_i.$$

Paso 9 Tome $w_{m-1} = z_{m-1}$.

Paso 10 Para $i = m-2, \ldots, 1$, tome $w_i = z_i - u_i w_{i+1}$.

Paso 11 SALIDA (t); (*Nota:* $t = t_j$.)
Para $i = 1, \ldots, m-1$, tome $x = ih$;
SALIDA (x, w_i). (*Nota:* $w_i = w_{i,j}$.)

Paso 12 PARAR. (*Procedimiento terminado.*) ■

EJEMPLO 3 Utilizaremos el método de Crank-Nicolson para aproximar la solución al problema de los ejemplos 1 y 2, que consiste en la ecuación

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < 1, \quad 0 < t,$$

sujeta a las condiciones

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$

y

$$u(x, 0) = \text{sen}\,(\pi x), \quad 0 \le x \le 1.$$

Las opciones $m = 10$, $h = 0.1$, $N = 50$, $k = 0.01$ y $\lambda = 1$ se usan en el algoritmo 12.3, como se hizo en los ejemplos anteriores. Los resultados de la tabla 12.5 indican el aumento

Tabla 12.5

x_i	$w_{i,50}$	$u(x_i, 0.5)$	$\lvert w_{i,50} - u(x_i, 0.5)\rvert$
0.0	0	0	
0.1	0.00230512	0.00222241	8.271×10^{-5}
0.2	0.00438461	0.00422728	1.573×10^{-4}
0.3	0.00603489	0.00581836	2.165×10^{-4}
0.4	0.00709444	0.00683989	2.546×10^{-4}
0.5	0.00745954	0.00719188	2.677×10^{-4}
0.6	0.00709444	0.00683989	2.546×10^{-4}
0.7	0.00603489	0.00581836	2.165×10^{-4}
0.8	0.00438461	0.00422728	1.573×10^{-4}
0.9	0.00230512	0.00222241	8.271×10^{-5}
1.0	0	0	

de exactitud del método de Crank-Nicolson respecto al de diferencias regresivas, o sea, la mejor de las técnicas explicadas con anterioridad. ■

CONJUNTO DE EJERCICIOS 12.2

1. Aproxime la solución a las siguientes ecuaciones diferenciales parciales usando el algoritmo de diferencias regresivas.

   a. $\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 2, 0 < t;$

   $$u(0, t) = u(2, t) = 0, \quad 0 < t,$$
   $$u(x, 0) = \text{sen}\, \frac{\pi}{2} x, \quad 0 \le x \le 2.$$

   Use $m = 4$, $T = 0.1$ y $N = 2$ y compare después sus respuestas con la solución real $u(x, t) = e^{-(\pi^2/4)t}\,\text{sen}\,\frac{\pi}{2}x$.

   b. $\dfrac{\partial u}{\partial t} - \frac{1}{16}\dfrac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 1, 0 < t;$

   $$u(0, t) = u(1, t) = 0, \quad 0 < t,$$
   $$u(x, 0) = 2\,\text{sen}\, 2\pi x, \quad 0 \le x \le 1.$$

   Use $m = 3$, $T = 0.1$ y $N = 2$ y compare después sus respuestas con la solución real $u(x, t) = 2e^{-(\pi^2/4)t}\,\text{sen}\, 2\pi x$.

2. Repita el ejercicio 1 usando el algoritmo de Crank-Nicolson.

3. Use el método de diferencias progresivas para aproximar la solución de las siguientes ecuaciones diferenciales parciales parabólicas.

   a. $\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 2, 0 < t;$

   $$u(0, t) = u(2, t) = 0, \quad 0 < t,$$
   $$u(x, 0) = \text{sen}\, 2\pi x, \quad 0 \le x \le 2.$$

   Use $h = 0.4$ y $k = 0.1$ y compare sus respuestas en $t = 0.5$ con la solución real $u(x, t) = e^{-4\pi^2 t}\,\text{sen}\, 2\pi x$. Después, use $h = 0.4$ y $k = 0.05$ y compare las respuestas.

   b. $\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < \pi, \quad 0 < t;$

   $$u(0, t) = u(\pi, t) = 0, \quad 0 < t,$$
   $$u(x, 0) = \text{sen}\, x, \quad 0 \le x \le \pi.$$

   Use $h = \pi/10$ y $k = 0.05$ y compare sus respuestas con la solución real $u(x, t) = e^{-t}\,\text{sen}\, x$ en $t = 0.5$.

   c. $\dfrac{\partial u}{\partial t} - \frac{4}{\pi^2}\dfrac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 4, \quad 0 < t;$

   $$u(0, t) = u(4, t) = 0, \quad 0 < t,$$
   $$u(x, 0) = \text{sen}\, \frac{\pi}{4} x\,(1 + 2\cos \frac{\pi}{4} x), \quad 0 \le x \le 4.$$

Use $h = 0.2$ y $k = 0.04$. Compare sus respuestas con la solución real $u(x, t) = e^{-t} \operatorname{sen} \frac{\pi}{2}x + e^{-t/4} \operatorname{sen} \frac{\pi}{4} x$ en $t = 0.4$.

d. $\dfrac{\partial u}{\partial t} - \dfrac{1}{\pi^2}\dfrac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 1, \quad 0 < t;$

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$
$$u(x, 0) = \cos \pi (x - \tfrac{1}{2}), \quad 0 \le x \le 1.$$

Use $h = 0.1$ y $k = 0.04$. Compare sus respuestas con la solución real $u(x, t) = e^{-t}\cos \pi (x - \frac{1}{2})$ en $t = 0.4$.

4. Repita el ejercicio 3 usando el algoritmo de las diferencias regresivas.

5. Repita el ejercicio 3 usando el algoritmo de Crank-Nicolson.

6. Repita el ejercicio 3 usando el método de Richardson.

7. Demuestre que los valores característicos de la matriz tridiagonal de $(m - 1)$ por $(m - 1)$ dados por

$$a_{ij} = \begin{cases} \lambda, & j = i - 1 \text{ o } j = i + 1, \\ 1 - 2\lambda, & j = i, \\ 0, & \text{de otro modo} \end{cases}$$

son

$$\mu_i = 1 - 4\lambda \left(\operatorname{sen} \frac{i\pi}{2m}\right)^2, \quad \text{para toda } i = 1, 2, \ldots, m - 1,$$

con los correspondientes vectores característicos $\mathbf{v}^{(i)}$, donde $v_j^{(i)} = \operatorname{sen} \frac{ij\pi}{m}$.

8. Demuestre que la matriz tridiagonal A de $(m - 1)$ por $(m - 1)$ dada por

$$a_{ij} = \begin{cases} -\lambda, & j = i - 1 \text{ o } j = i + 1, \\ 1 + 2\lambda, & j = i, \\ 0, & \text{de otro modo,} \end{cases}$$

donde $\lambda > 0$ es definida positiva, diagonalmente dominante y tiene los valores característicos

$$\mu_i = 1 + 4\lambda \left(\operatorname{sen} \frac{i\pi}{2m}\right)^2, \quad \text{para toda } i = 1, 2, \ldots, m - 1,$$

con los correspondientes vectores característicos $\mathbf{v}^{(i)}$, donde $v_j^{(i)} = \operatorname{sen} \frac{ij\pi}{m}$.

9. Modifique los algoritmos 12.2 y 12.3 para incluir la ecuación diferencial parcial parabólica

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = F(x), \quad 0 < x < l, \quad 0 < t;$$

$$u(0, t) = u(l, t) = 0, \quad 0 < t,$$
$$u(x, 0) = f(x), \quad 0 \le x \le l.$$

10. Use los resultados del ejercicio 9 para aproximar la solución de

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 2, \quad 0 < x < 1, \quad 0 < t;$$

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$
$$u(x, 0) = \operatorname{sen} \pi x + x(1 - x),$$

con $h = 0.1$ y $k = 0.01$. Compare su respuesta con la solución real $u(x, t) = e^{-\pi^2 t}$ sen $\pi x + x(1 - x)$ en $t = 0.25$.

11. Modifique los algoritmos 12.2 y 12.3 para incluir la ecuación diferencial parcial

$$\frac{\partial u}{\partial t} - \alpha^2 \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < l, \quad 0 < t;$$

$$u(0, t) = \phi(t), \ u(l, t) = \Psi(t), \quad 0 < t,$$

$$u(x, 0) = f(x), \quad 0 \leq x \leq l,$$

donde $f(0) = \phi(0)$ y $f(l) = \Psi(0)$.

12. La temperatura $u(x, t)$ de una varilla larga y delgada, de sección transversal constante y de un material conductor homogéneo está regida por la ecuación unidimensional de calor. Si se genera calor en el material (por ejemplo, debido a la resistencia a la corriente o a la reacción nuclear), la ecuación se convierte en

$$\frac{\partial^2 u}{\partial x^2} + \frac{Kr}{\rho C} = K\frac{\partial u}{\partial t}, \quad 0 < x < l, \quad 0 < t,$$

donde l es la longitud, ρ es la densidad, C es el calor específico y K es la difusividad térmica de la varilla. La función $r = r(x, t, u)$ representa el calor generado por unidad de volumen. Suponga que

$$l = 1.5 \text{ cm}, \qquad K = 1.04 \text{ cal/cm} \cdot \text{deg} \cdot \text{s},$$
$$\rho = 10.6 \text{ g/cm}^3, \qquad C = 0.056 \text{ cal/g} \cdot \text{deg},$$

y que

$$r(x, t, u) = 5.0 \text{ cal/cm}^3 \cdot \text{s}.$$

Si los extremos de la varilla se mantienen en 0°C, entonces

$$u(0, t) = u(l, t) = 0, \quad t > 0.$$

Suponga que la distribución inicial de la temperatura está dada por

$$u(x, 0) = \text{sen}\, \frac{\pi x}{l}, \quad 0 \leq x \leq l.$$

Use los resultados del ejercicio 9 para aproximar la distribución de la temperatura con $h = 0.15$ y con $k = 0.0225$.

13. Sagar y Payne [SP] analizan las relaciones de esfuerzo-deformación y las propiedades materiales de un cilindro sujeto alternativamente al calentamiento y al enfriamiento, y consideran la ecuación

$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} = \frac{1}{4K}\frac{\partial T}{\partial t}, \quad \frac{1}{2} < r < 1, 0 < T,$$

donde $T = T(r, t)$ es la temperatura, r es la distancia radial respecto al centro del cilindro, t es el tiempo y K es el coeficiente de difusividad.

a. Obtenga las aproximaciones a $T(r, 10)$ para un cilindro con radio externo 1, dadas las condiciones iniciales y de frontera:

$$T(1, t) = 100 + 40t, \quad 0 \leq t \leq 10;$$

$$T\left(\frac{1}{2}, t\right) = t, \quad 0 \leq t \leq 10;$$

$$T(r, 0) = 200(r - 0.5), \quad 0.5 \leq r \leq 1.$$

Use una modificación del método de diferencias regresivas con $K = 0.1$, $k = 0.5$ y $h = \Delta r = 0.1$.

b. Por medio de la distribución de la temperatura del inciso (a), calcule la deformación I aproximando la integral

$$I = \int_{0.5}^{1} \alpha T(r, t) r \, dr,$$

donde $\alpha = 10.7$ y $t = 10$. Use el método del trapecio compuesto con $n = 5$.

12.3 Ecuaciones diferenciales parciales hiperbólicas

En esta sección estudiaremos la solución numérica para la **ecuación de onda**, que es un ejemplo de una ecuación diferencial parcial *hiperbólica*. La ecuación de onda está dada por la ecuación diferencial

$$\frac{\partial^2 u}{\partial t^2}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < l, \quad t > 0, \tag{12.16}$$

sujeta a las condiciones

$$u(0, t) = u(l, t) = 0, \quad \text{para} \quad t > 0,$$

$$u(x, 0) = f(x), \quad \text{y} \quad \frac{\partial u}{\partial x}(x, 0) = g(x), \quad \text{para} \quad 0 \le x \le l,$$

donde α es una constante. Para establecer el método de diferencias finitas, se selecciona un entero $m > 0$ y el tamaño de paso de tiempo $k > 0$. Con $h = l/m$ los puntos de red (x_i, t_j) son

$$x_i = ih, \quad \text{y} \quad t_j = jk,$$

para cada $i = 0, 1, \ldots, m$ y $j = 0, 1, \ldots$ En cualquier punto de red interior (x_i, t_j) la ecuación de onda se transforma en

$$\frac{\partial^2 u}{\partial t^2}(x_i, t_j) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x_i, t_j) = 0. \tag{12.17}$$

El método de diferencias se obtiene usando el cociente de diferencias centradas en las segundas derivadas parciales dadas por

$$\frac{\partial^2 u}{\partial t^2}(x_i, t_j) = \frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} - \frac{k^2}{12}\frac{\partial^4 u}{\partial t^4}(x_i, \mu_j),$$

donde $\mu_j \in (t_{j-1}, t_{j+1})$ y

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) = \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, t_j),$$

donde $\xi_i \in (x_{i-1}, x_{i+1})$. Al sustituir estas expresiones en la ecuación (12.17), obtenemos

$$\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} - \alpha^2 \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$

$$= \frac{1}{12}\left[k^2 \frac{\partial^4 u}{\partial t^4}(x_i, \mu_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j)\right].$$

Si ignoramos el término de error

$$\tau_{i,j} = \frac{1}{12}\left[k^2 \frac{\partial^4 u}{\partial t^4}(x_i, \mu_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j)\right],$$

obtenemos la ecuación de diferencias

$$\frac{w_{i,j+1} - 2w_{i,j} + w_{i,j-1}}{k^2} - \alpha^2 \frac{w_{i+1,j} - 2w_{i,j} + w_{i-1,j}}{h^2} = 0.$$

Si $\lambda = \alpha k/h$, podemos escribir la ecuación de diferencias como

$$w_{i,j+1} - 2w_{i,j} + w_{i,j-1} - \lambda^2 w_{i+1,j} + 2\lambda^2 w_{i,j} - \lambda^2 w_{i-1,j} = 0$$

y resolver para $w_{i,j+1}$, o sea, la aproximación más avanzada del paso de tiempo, para obtener

$$w_{i,j+1} = 2(1 - \lambda^2)w_{i,j} + \lambda^2(w_{i+1,j} + w_{i-1,j}) - w_{i,j-1}. \tag{12.18}$$

Esta ecuación es aplicable para toda $i = 1, 2, \ldots, m-1$ y $j = 1, 2, \ldots$ Las condiciones de frontera nos dan

$$w_{0,j} = w_{m,j} = 0, \text{ para cada } j = 1, 2, 3, \ldots \tag{12.19}$$

y la condición inicial implica que

$$w_{i,0} = f(x_i), \text{ para cada } i = 1, 2, \ldots, m-1. \tag{12.20}$$

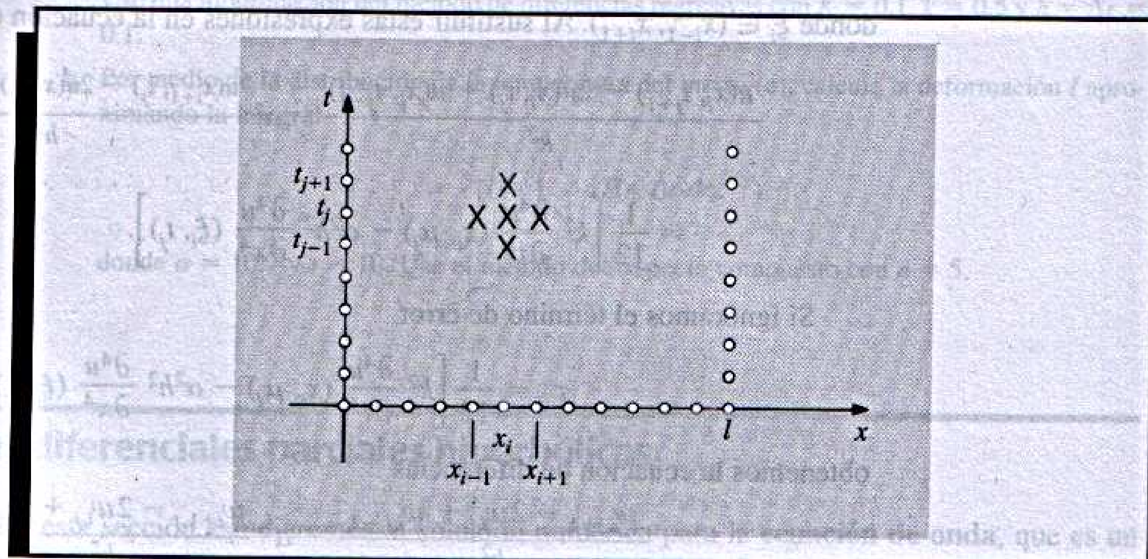
Al escribir este conjunto de ecuaciones en forma matricial, obtenemos

$$\begin{bmatrix} w_{1,j+1} \\ w_{2,j+1} \\ \vdots \\ w_{m-1,j+1} \end{bmatrix} = \begin{bmatrix} 2(1-\lambda^2) & \lambda^2 & 0 & \cdots & 0 \\ \lambda^2 & 2(1-\lambda^2) & \lambda^2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \lambda^2 \\ 0 & \cdots & 0 & \lambda^2 & 2(1-\lambda^2) \end{bmatrix} \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ w_{m-1,j} \end{bmatrix} - \begin{bmatrix} w_{1,j-1} \\ w_{2,j-1} \\ \vdots \\ w_{m-1,j-1} \end{bmatrix}. \tag{12.21}$$

Las ecuaciones (12.18) y (12.19) implican que el $(j+1)$-ésimo paso de tiempo requiere valores de los j-ésimo y $(j-1)$-ésimo pasos. (Véase la Fig. 12.10.) Esto produce un pequeño problema inicial, porque los valores de $j = 0$ están dados por la ecuación (12.20), pero los valores de $j = 1$, que se necesitan en la ecuación (12.18) para calcular $w_{i,2}$, deben obtenerse de la condición de velocidad inicial

$$\frac{\partial u}{\partial t}(x, 0) = g(x), \quad 0 \le x \le l.$$

Figura 12.10



Un procedimiento consiste en reemplazar $\partial u/\partial t$ por una aproximación de diferencias progresivas,

$$\frac{\partial u}{\partial t}(x_i, 0) = \frac{u(x_i, t_1) - u(x_i, 0)}{k} - \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \tilde{\mu}_i), \tag{12.22}$$

para cierta $\tilde{\mu}_i$ en $(0, t_1)$. Al resolver para $u(x_i, t_1)$ obtenemos

$$u(x_i, t_1) = u(x_i, 0) + k\frac{\partial u}{\partial t}(x_i, 0) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \tilde{\mu}_i)$$

$$= u(x_i, 0) + kg(x_i) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \tilde{\mu}_i).$$

En consecuencia,

$$w_{i,1} = w_{i,0} + kg(x_i), \quad \text{para cada } i = 1, \ldots, m-1. \tag{12.23}$$

Sin embargo, esto da una aproximación con un error de sólo $O(k)$. Podemos obtener una mejor aproximación a $u(x_i, 0)$. Considere la ecuación

$$u(x_i, t_1) = u(x_i, 0) + k\frac{\partial u}{\partial t}(x_i, 0) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, 0) + \frac{k^3}{6}\frac{\partial^3 u}{\partial t^3}(x_i, \hat{\mu}_i),$$

para cierta $\hat{\mu}_i$ en $(0, t_1)$, que proviene de desarrollar $u(x_i, t_1)$ con el segundo polinomio de Maclaurin en t. Si f'' existe, entonces

$$\frac{\partial^2 u}{\partial t^2}(x_i, 0) = \alpha^2\frac{\partial^2 u}{\partial x^2}(x_i, 0) = \alpha^2 + \frac{d^2 f}{dx^2}(x_i) = \alpha^2 f''(x_i)$$

y

$$u(x_i, t_1) = u(x_i, 0) + kg(x_i) + \frac{\alpha^2 k^2}{2} f''(x_i) + \frac{k^3}{6}\frac{\partial^3 u}{\partial t^3}(x_i, \hat{\mu}_i),$$

lo que produce una aproximación con error $O(k^3)$:

$$w_{i1} = w_{i0} + kg(x_i) + \frac{\alpha^2 k^2}{2} f''(x_i).$$

Si $f \in C^4[0, 1]$ pero no disponemos de $f''(x_i)$, podemos usar la ecuación en diferencias de (4.9) para escribir

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} - \frac{h^2}{12} f^{(4)}(\tilde{\xi}_i),$$

para alguna $\tilde{\xi}_i$ en (x_{i-1}, x_{i+1}). Esto implica que

$$u(x_i, t_1) = u(x_i, 0) + kg(x_i) + \frac{k^2\alpha^2}{2h^2}[f(x_{i+1}) - 2f(x_i) + f(x_{i-1})] + O(k^3 + h^2k^2).$$

Si $\lambda = (k\alpha/h)$, entonces

$$\begin{aligned} u(x_i, t_1) &= u(x_i, 0) + kg(x_i) + \frac{\lambda^2}{2}[f(x_{i+1}) - 2f(x_i) + f(x_{i-1})] + O(k^3 + h^2k^2) \\ &= (1 - \lambda^2) f(x_i) + \frac{\lambda^2}{2} f(x_{i+1}) + \frac{\lambda^2}{2} f(x_{i-1}) + kg(x_i) + O(k^3 + h^2k^2). \end{aligned}$$

Así, podemos usar la ecuación en diferencias

$$w_{i,1} = (1 - \lambda^2) f(x_i) + \frac{\lambda^2}{2} f(x_{i+1}) + \frac{\lambda^2}{2} f(x_{i-1}) + kg(x_i), \tag{12.24}$$

para calcular $w_{i,1}$, para cada $i = 1, 2, \ldots, m - 1$.

El algoritmo 12.4 usa la ecuación (12.24) para aproximar $w_{i,1}$, aunque la ecuación (12.23) también sirve. Se supone que existe una cota superior para el valor de t que se usa en la técnica de paro, y que $k = T/N$, donde también N está dado.

ALGORITMO 12.4

Método de diferencias finitas para la ecuación de onda

Para aproximar la solución de la ecuación de onda

$$\frac{\partial^2 u}{\partial t^2}(x, t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < l, \quad 0 < t < T,$$

sujeta a las condiciones de frontera

$$u(0, t) = u(l, t) = 0, \; 0 < t < T,$$

y a las condiciones iniciales

$$u(x, 0) = f(x), \quad 0 \le x \le l,$$

$$\frac{\partial u}{\partial t}(x, 0) = g(x), \quad 0 \le x \le l:$$

ENTRADA extremo l; tiempo máximo T; constante α; enteros $m \ge 2$, $N \ge 2$.

SALIDA aproximaciones $w_{i,j}$ a $u(x_i, t_j)$ para toda $i = 0, \ldots, m$ y $j = 0, \ldots, N$.

Paso 1 Tome $h = l/m$;
$k = T/N$;
$\lambda = ka/h$.

Paso 2 Para $j = 1, \ldots, N$, tome $w_{0,j} = 0$;
$w_{m,j} = 0$.

Paso 3 Tome $w_{0,0} = f(0)$;
$w_{m,0} = f(l)$.

Paso 4 Para $i = 1, \ldots, m-1$ *(Inicialice para $t = 0$ y $t = k$.)*
tome $w_{i,0} = f(ih)$;

$$w_{i,1} = (1 - \lambda^2) f(ih) + \frac{\lambda^2}{2} [f((i+1)h) + f((i-1)h)] + kg(ih).$$

Paso 5 Para $j = 1, \ldots, N-1$ *(Realice una multiplicación de matrices.)*
para $i = 1, \ldots, m-1$
tome $w_{i,j+1} = 2(1 - \lambda^2) w_{i,j} + \lambda^2 (w_{i+1,j} + w_{i-1,j}) + w_{i,j-1}$.

Paso 6 Para $j = 0, \ldots, N$
tome $t = jk$;
para $i = 0, \ldots, m$
tome $x = ih$;
SALIDA $(x, t, w_{i,j})$.

Paso 7 PARAR. *(Procedimiento terminado.)* ■

EJEMPLO 1 Considérese el problema hiperbólico

$$\frac{\partial^2 u}{\partial t^2}(x, t) - 4\,\frac{\partial^2 u}{\partial x^2}(x, t) = 0, \quad 0 < x < 1, \quad 0 < t,$$

con las condiciones de frontera

$$u(0, t) = u(1, t) = 0, \quad \text{para} \quad 0 < t,$$

y con las condiciones iniciales

$$u(x, 0) = \operatorname{sen}(\pi x), \quad 0 \le x \le 1, \quad \text{y} \quad \frac{\partial u}{\partial t}(x, 0) = 0, \quad 0 \le x \le 1.$$

Se puede verificar fácilmente que la solución a este problema es

$$u(x, t) = \operatorname{sen} \pi x \cos 2\pi t.$$

En este ejemplo se emplea el algoritmo de diferencias finitas 12.4 con $m = 10$, $T = 1$ y $N = 20$, lo cual significa que $h = 0.1$, $k = 0.05$ y $\lambda = 1$. La tabla 12.6 contiene los resultados de la aproximación $w_{i,N}$ para $i = 0, 1, \ldots, 10$. Los valores que se dan en la tabla son correctos en las cifras decimales dadas. ■

Tabla 12.6

x_i	$w_{i,20}$
0.0	0.0000000000
0.1	0.3090169944
0.2	0.5877852523
0.3	0.8090169944
0.4	0.9510565163
0.5	1.0000000000
0.6	0.9510565163
0.7	0.8090169944
0.8	0.5877852523
0.9	0.3090169944
1.0	0.0000000000

Los resultados del ejemplo fueron muy exactos, más de lo que el error de truncamiento de $O(k^2 + h^2)$ nos haría suponer. La explicación de ello se encuentra en el hecho de que la solución verdadera de la ecuación es infinitamente diferenciable. Cuando es así, al utilizar la serie de Taylor obtenemos

$$\frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$
$$= \frac{\partial^2 u}{\partial x^2}(x_i, t_j) + 2\left[\frac{h^2}{4!}\,\frac{\partial^4 u}{\partial x^4}(x_i, t_j) + \frac{h^4}{6!}\,\frac{\partial^6 u}{\partial x^6}(x_i, t_j) + \ldots\right]$$

y

$$\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2}$$
$$= \frac{\partial^2 u}{\partial t^2}(x_i, t_j) + 2\left[\frac{k^2}{4!}\,\frac{\partial^4 u}{\partial t^4}(x_i, t_j) + \frac{h^4}{6!}\,\frac{\partial^6 u}{\partial t^6}(x_i, t_j) + \ldots\right].$$

Puesto que $u(x, t)$ satisface la ecuación diferencial parcial,

$$\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{k^2} - \alpha^2\frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2}$$
$$= 2\left[\frac{1}{4!}\left(k^2\,\frac{\partial^4 u}{\partial t^4}(x_i, t_j) - \alpha^2 h^2\,\frac{\partial^4 u}{\partial x^4}(x_i, t_j)\right)\right.$$
$$\left. + \frac{1}{6!}\left(k^4\,\frac{\partial^6 u}{\partial t^6}(x_i, t_j) - \alpha^2 h^4\,\frac{\partial^6 u}{\partial x^6}(x_i, t_j)\right) + \ldots\right]. \quad (12.25)$$

Sin embargo, al derivar la ecuación de onda,

$$k^2\,\frac{\partial^4 u}{\partial t^4}(x_i, t_j) = k^2\,\frac{\partial^2}{\partial t^2}\left[\alpha^2\frac{\partial^2 u}{\partial x^2}(x_i, t_j)\right] = \alpha^2 k^2\frac{\partial^2}{\partial x^2}\left[\frac{\partial^2 u}{\partial t^2}(x_i, t_j)\right]$$
$$= \alpha^2 k^2\frac{\partial^2}{\partial x^2}\left[\alpha^2\frac{\partial^2 u}{\partial x^2}(x_i, t_j)\right] = \alpha^4 k^2\frac{\partial^4 u}{\partial x^4}(x_i, t_j).$$

y vemos que, como $\lambda^2 = (\alpha^2 k^2/h^2) = 1$, tenemos

$$\frac{1}{4!}\left[k^2 \frac{\partial^4 u}{\partial t^4}(x_i, t_j) - \alpha^2 h^2 \frac{\partial^4 u}{\partial x^4}(x_i, t_j)\right] = \frac{\alpha^2}{4!}\,[\alpha^2 k^2 - h^2]\,\frac{\partial^4 u}{\partial x^4}(x_i, t_j) = 0.$$

Continuando de esta forma, todos los términos del lado derecho de (12.25) son 0, lo cual supone un error de truncamiento local 0. Los únicos errores del ejemplo 1 son los debidos a la aproximación de $w_{i,1}$ y al redondeo.

Al igual que en el caso del método de diferencias progresivas para la ecuación del calor, el método de las diferencias finitas explícitas para la ecuación de onda también presenta problemas de estabilidad. De hecho, para que el método sea estable es necesario que $\lambda = \alpha k/h \leq 1$. (Véase [IK, p. 489].) El método explícito que se da en el algoritmo 12.4, con $\lambda \leq 1$, es de $O(h^2 + k^2)$ convergente si f y g son suficientemente diferenciables. Para una comprobación de lo anterior consulte [IK, p. 491].

Aunque no los describiremos aquí, existen métodos implícitos incondicionalmente estables, y una explicación de ellos la puede encontrar en [Am, p. 199], en [Mi] y en [Sm,B].

CONJUNTO DE EJERCICIOS 12.3

1. Aproxime la solución de la ecuación de onda

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 1, \quad 0 < t;$$
$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$
$$u(x, 0) = \text{sen}\,\pi x, \quad 0 \leq x \leq 1,$$
$$\frac{\partial u}{\partial t}(x, 0) = 0, \quad 0 \leq x \leq 1,$$

usando el algoritmo de diferencias finitas con $m = 4$, $N = 4$ y $T = 1.0$. Compare después los resultados con la solución real $u(x, t) = \cos \pi t \,\text{sen}\, \pi x$ en $t = 1.0$.

2. Aproxime la solución de la ecuación de onda

$$\frac{\partial^2 u}{\partial t^2} - \frac{1}{16\pi^2}\frac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 0.5, \quad 0 < t;$$
$$u(0, t) = u(0.5, t) = 0, \quad 0 < t,$$
$$u(x, 0) = 0, \quad 0 \leq x \leq 0.5,$$
$$\frac{\partial u}{\partial t}(x, 0) = \text{sen}\, 4\pi x, \quad 0 \leq x \leq 0.5,$$

usando el algoritmo de diferencias finitas con $m = 4$, $N = 4$ y $T = 0.5$, y compare sus resultados con la solución real $u(x, t)$ sen t sen $4\pi x$ en $t = 0.5$.

3. Aproxime la solución de la ecuación de onda

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < \pi, \quad 0 < t;$$
$$u(0, t) = u(\pi, t) = 0, \quad 0 < t,$$

$$u(x, 0) = \operatorname{sen} x, \quad 0 \leq x \leq \pi,$$

$$\frac{\partial u}{\partial t}(x, 0) = 0, \quad 0 \leq x \leq \pi,$$

cuando el algoritmo de diferencias finitas con $h = \pi/10$ y $k = 0.05$, con $h = \pi/20$ y con $k = 0.1$ y luego con $h = \pi/20$ y con $k = 0.05$. Compare después los resultados con la solución real $u(x, t) = \cos t \operatorname{sen} x$ en $t = 0.5$.

4. Repita el ejercicio 3 usando en el paso 4 del algoritmo 12.4 la aproximación

$$w_{i,1} = w_{i,0} + kg(x_i), \quad \text{para cada } i = 1, \ldots, m - 1.$$

5. Aproxime la solución de la ecuación de onda

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 1, 0 < t;$$

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$

$$u(x, 0) = \operatorname{sen} 2\pi x, \quad 0 \leq x \leq 1,$$

$$\frac{\partial u}{\partial t}(x, 0) = 2\pi \operatorname{sen} 2\pi x, \quad 0 \leq x \leq 1,$$

usando el algoritmo 12.4 con $h = 0.1$ y con $k = 0.1$. Compare sus resultados con la solución real $u(x, t) = \operatorname{sen} 2\pi x(\cos 2\pi t + \operatorname{sen} 2\pi t)$, en $t = 0.3$.

6. Aproxime la solución de la ecuación de onda en $t = 0.5$

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 < x < 1, \quad 0 < t;$$

$$u(0, t) = u(1, t) = 0, \quad 0 < t,$$

$$u(x, 0) = \begin{cases} 1, & 0 \leq x \leq \frac{1}{2} \\ -1, & \frac{1}{2} < x \leq 1, \end{cases}$$

$$\frac{\partial u}{\partial t}(x, 0) = 0, \quad 0 \leq x \leq 1.$$

usando el algoritmo 12.4 con $h = 0.1$ y $k = 0.1$.

7. En un tubo de órgano, la presión del aire $p(x, t)$ se rige por la ecuación de onda

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}, \quad 0 < x < l, 0 < t,$$

donde l es la longitud del tubo y c es una constante física. Si el tubo se encuentra abierto, las condiciones de frontera estarán dadas por

$$p(0, t) = p_0 \quad \text{y} \quad p(l, t) = p_0.$$

Si el tubo está cerrado en el extremo donde $x = l$, las condiciones de frontera serán

$$p(0, t) = p_0 \quad \text{y} \quad \frac{\partial p}{\partial x}(l, t) = 0.$$

Suponga que $c = 1$, $l = 1$ y que las condiciones iniciales son

$$p(x, 0) = p_0 \cos 2\pi x, \quad \text{y} \quad \frac{\partial p}{\partial t}(x, 0) = 0, \quad 0 \leq x \leq 1.$$

a. Aproxime la presión de un tubo abierto con $p_0 = 0.9$ en $x = \frac{1}{2}$ para $t = 0.5$ y $t = 1$, usando el algoritmo 12.4 con $h = k = 0.1$.

b. Modifique el algoritmo 12.4 para el problema del tubo de órgano cerrado con $p_0 = 0.9$ y luego aproxime $p(0.5, 0.5)$ y $p(0.5, 1)$ usando $h = k = 0.1$.

8. En una línea de transmisión eléctrica de longitud l, que conduce una corriente alterna de alta frecuencia (llamada línea "sin pérdida"), el voltaje V y la corriente i se describe por medio de

$$\frac{\partial^2 V}{\partial x^2} = LC\frac{\partial^2 V}{\partial t^2}, \quad 0 < x < l, 0 < t;$$

$$\frac{\partial^2 i}{\partial x^2} = LC\frac{\partial^2 i}{\partial t^2}, \quad 0 < x < l, 0 < t;$$

donde L es la inductancia por longitud unitaria y C es la capacitancia por longitud unitaria. Suponga que la línea tiene 200 pies de largo y que las constantes C y L están dadas por

$$C = 0.1 \text{ farads/pies} \quad \text{y} \quad L = 0.3 \text{ henries/pies}.$$

Suponga, además, que el voltaje y la corriente también satisfacen

$$V(0, t) = V(200, t) = 0, \quad 0 < t;$$
$$V(x, 0) = 110 \operatorname{sen}\frac{\pi x}{200}, \quad 0 \le x \le 200;$$
$$\frac{\partial V}{\partial t}(x, 0) = 0, \quad 0 \le x \le 200;$$
$$i(0, t) = i(200, t) = 0, \; 0 < t;$$
$$i(x, 0) = 5.5 \cos\frac{\pi x}{200}, \; 0 \le x \le 200;$$

y

$$\frac{\partial i}{\partial t}(x, 0) = 0, \quad 0 \le x \le 200.$$

Aproxime el voltaje y la corriente en $t = 0.2$ y $t = 0.5$ usando el algoritmo 12.4 con $h = 10$ y $k = 0.1$.

12.4 Una introducción al método de elementos finitos

El **método de elementos finitos** es muy parecido al de Rayleigh-Ritz, que explicamos en la sección 11.5 para aproximar la solución a los problemas de valor de frontera de dos puntos. Originalmente se ideó para utilizarse en ingeniería civil, pero hoy sirve para aproximar las soluciones de las ecuaciones diferenciales parciales que se presentan en todos los campos de las matemáticas aplicadas.

Una ventaja de este método sobre los de diferencias finitas es la facilidad relativa con que se manejan las condiciones de frontera del problema. Muchos problemas físicos tienen este tipo de condiciones que incluyen derivadas y fronteras de forma irregular. Resulta difícil manejar esta clase de condiciones de frontera con los métodos de diferencias finitas, dado que cada condición que contenga una derivada debe aproximarse mediante un cociente de diferencias en los puntos de red, y la forma irregular de la frontera dificulta la

colocación de los puntos de red. El método de elementos finitos incluye las condiciones de frontera como integrales en una funcional que va a reducirse al mínimo, de modo que el procedimiento de construcción es independiente de las condiciones particulares de frontera del problema.

En nuestra exposición consideraremos la ecuación diferencial parcial

$$\frac{\partial}{\partial x}\left(p(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(q(x, y)\frac{\partial u}{\partial y}\right) + r(x, y)u(x, y) = f(x, y), \tag{12.26}$$

con $(x, y) \in \mathcal{D}$ donde $\mathcal{D}$ es una región plana con fronteras $\mathcal{S}$.

Las condiciones de frontera de la forma

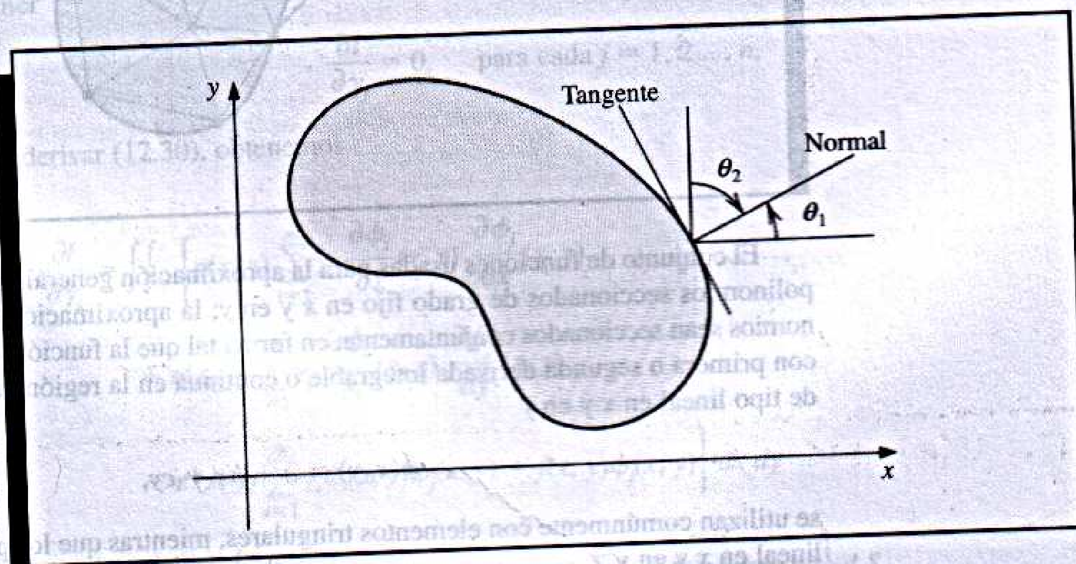
$$u(x, y) = g(x, y) \tag{12.27}$$

se imponen en una parte $\mathcal{S}_1$ de la frontera. En el resto de la frontera, $\mathcal{S}_2$, se requiere que la solución $u(x, y)$ satisfaga

$$p(x, y)\frac{\partial u}{\partial x}(x, y)\cos\theta_1 + q(x, y)\frac{\partial u}{\partial y}(x, y)\cos\theta_2 + g_1(x, y)u(x, y) = g_2(x, y), \tag{12.28}$$

donde θ_1 y θ_2 son los ángulos de dirección de la normal hacia afuera respecto a la frontera en el punto (x, y). (Véase la Fig. 12.11.)

Figura 12.11



Los problemas físicos de la mecánica de sólidos y de elasticidad tienen ecuaciones diferenciales parciales semejantes a la ecuación (12.26). La solución de este tipo de problemas normalmente reduce al mínimo cierto funcional, que involucra integrales, de una clase de funciones determinadas por el problema.

Supongamos que p, q, r y f son continuas en $\mathcal{D} \cup \mathcal{S}$, que p y q tienen primeras derivadas parciales continuas y que g_1 y g_2 son continuas en $\mathcal{S}_2$. Supongamos, además, que $p(x, y) > 0$, $q(x, y) > 0$, $r(x, y) \leq 0$ y $g_1(x, y) > 0$. Entonces, una solución de la ecuación (12.26) minimiza en forma única el funcional

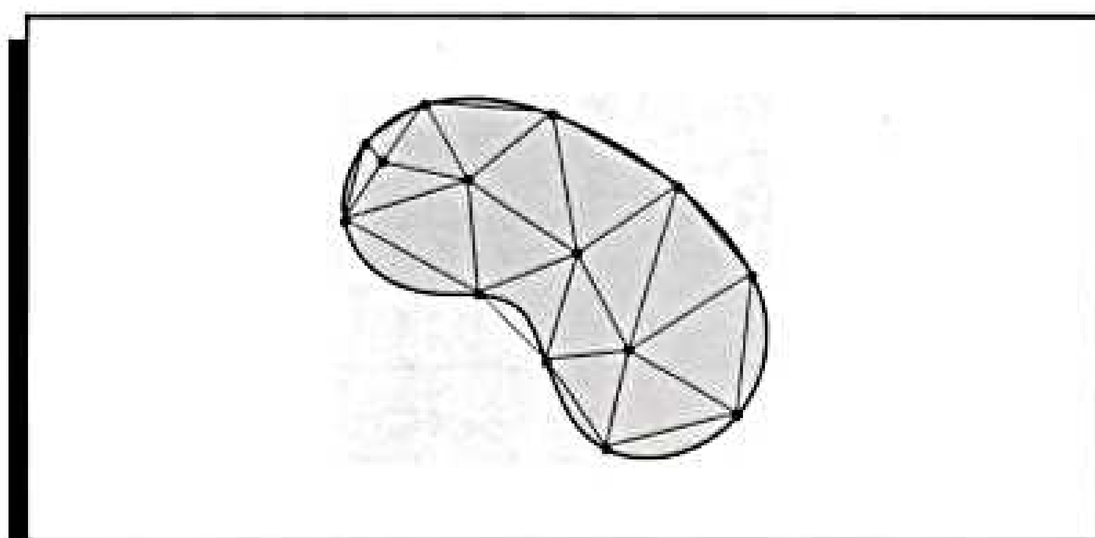
$$I[w] = \iint_{\mathcal{D}} \left\{ \frac{1}{2} \left[p(x, y) \left(\frac{\partial w}{\partial x} \right)^2 + q(x, y) \left(\frac{\partial w}{\partial y} \right)^2 - r(x, y) w^2 \right] + f(x, y) w \right\} dx\, dy$$

$$+ \int_{\mathcal{S}_2} \left\{ -g_2(x, y) w + \frac{1}{2} g_1(x, y) w^2 \right\} dS \qquad (12.29)$$

sobre todas las funciones w que satisfacen la ecuación (12.27) en $\mathcal{S}_1$ que son continuamente diferenciables dos veces. El método de los elementos finitos aproxima esta solución al reducir al mínimo el funcional I en una clase más pequeña de funciones, como lo hizo el método de Rayleigh-Ritz en el problema con valor de frontera considerado en la sección 11.5.

El primer paso consiste en dividir la región en una cantidad finita de secciones o elementos de tamaño regular, ya sea rectángulos o triángulos. (Véase la Fig. 12.12.)

Figura 12.12



El conjunto de funciones usadas para la aproximación generalmente es un conjunto de polinomios seccionados de grado fijo en x y en y; la aproximación requiere que los polinomios sean seccionados conjuntamente, en forma tal que la función resultante sea continua con primera o segunda derivada integrable o continua en la región entera. Los polinomios de tipo lineal en x y en y

$$\phi(x, y) = a + bx + cy,$$

se utilizan comúnmente con elementos tringulares, mientras que los polinomios de tipo bilineal en x y en y,

$$\phi(x, y) = a + bx + cy + dxy,$$

se usan con elementos rectangulares.

Para efectos de nuestro análisis, supondremos que la región $\mathcal{D}$ se subdividió en elementos triangulares. El conjunto de triángulos se denota con D, y a sus vértices se les da el nombre de **nodos**. El método busca una aproximación de la forma

$$\phi(x, y) = \sum_{i=1}^{m} \gamma_i \phi_i(x, y),$$

donde $\phi_1, \phi_2, \ldots, \phi_m$ son polinomios lineales seccionados linealmente independientes y $\gamma_1, \gamma_2, \ldots, \gamma_m$ son constantes. Algunas de estas constantes, digamos, $\gamma_{n+1}, \gamma_{n+2}, \ldots, \gamma_m$, sirven para asegurarse de que la condición de frontera,

$$\phi(x, y) = g(x, y),$$

se satisfaga en $\mathcal{S}_1$, y las constantes restantes $\gamma_1, \gamma_2, \ldots, \gamma_n$ se emplean para minimizar el funcional $I[\sum_{i=1}^{m} \gamma_i\phi_i]$.

De acuerdo con la ecuación (12.29), el funcional tiene la forma

$$\begin{aligned} I[\phi] &= I\left[\sum_{i=1}^{m} \gamma_i\phi_i\right] \\ &= \iint_{\mathcal{D}} \left(\frac{1}{2}\left\{p(x, y)\left[\sum_{i=1}^{m} \gamma_i \frac{\partial \phi_i}{\partial x}(x, y)\right]^2 + q(x, y)\left[\sum_{i=1}^{m} \gamma_i \frac{\partial \phi_i}{\partial y}(x, y)\right]^2\right.\right. \\ &\quad \left.\left. - r(x, y)\left[\sum_{i=1}^{m} \gamma_i\phi_i(x, y)\right]^2\right\} + f(x, y)\sum_{i=1}^{m} \gamma_i\phi_i(x, y)\right) dy\, dx \\ &\quad + \int_{\mathcal{S}_2}\left\{- g_2(x, y)\sum_{i=1}^{m} \gamma_i\phi_i(x, y) + \frac{1}{2} g_1(x, y)\left[\sum_{i=1}^{m} \gamma_i\phi_i(x, y)\right]^2\right\} dS. \end{aligned} \tag{12.30}$$

Para que ocurra un mínimo, considerando I como función de $\gamma_1, \gamma_2, \ldots, \gamma_n$, es necesario tener

$$\frac{\partial I}{\partial \gamma_j} = 0, \quad \text{para cada } j = 1, 2, \ldots, n.$$

Al derivar (12.30), obtenemos

$$\begin{aligned} \frac{\partial I}{\partial \gamma_j} &= \iint_{\mathcal{D}} \left\{p(x, y)\sum_{i=1}^{m}\gamma_i \frac{\partial \phi_i}{\partial x}(x, y)\, \frac{\partial \phi_j}{\partial x}(x, y)\right. \\ &\quad + q(x, y)\sum_{i=1}^{m}\gamma_i \frac{\partial \phi_i}{\partial y}(x, y)\, \frac{\partial \phi_j}{\partial y}(x, y) \\ &\quad \left. - r(x, y)\sum_{i=1}^{m}\gamma_i\phi_i(x, y)\phi_j(x, y) + f(x, y)\phi_j(x, y)\right\} dx\, dy \\ &\quad + \int_{\mathcal{S}_2}\left\{- g_2(x, y)\phi_j(x, y) + g_1(x, y)\sum_{i=1}^{m}\gamma_i\phi_i(x, y)\phi_j(x, y)\right\} dS, \end{aligned}$$

por lo que

$$\begin{aligned} 0 = \sum_{i=1}^{m}\Bigg[\iint_{\mathcal{D}} &\left\{p(x, y)\frac{\partial \phi_i}{\partial x}(x, y)\, \frac{\partial \phi_j}{\partial x}(x, y) + q(x, y)\, \frac{\partial \phi_i}{\partial y}(x, y)\, \frac{\partial \phi_j}{\partial y}(x, y)\right. \\ &\left. - r(x, y)\phi_i(x, y)\phi_j(x,y)\right\} dx\, dy \end{aligned}$$

$$+ \int_{\mathcal{S}_2} g_1(x, y)\phi_i(x, y)\phi_j(x, y)\, dS\Bigg]\gamma_i$$

$$+ \iint_{\mathcal{D}} f(x, y)\phi_j(x, y)\, dx\, dy - \int_{\mathcal{S}_2} g_2(x, y)\phi_j(x, y)\, dS,$$

para toda $j = 1, 2, \ldots, n$. Este conjunto de ecuaciones puede escribirse como un sistema lineal:

$$A\mathbf{c} = \mathbf{b},$$

donde $\mathbf{c} = (\gamma_1, \ldots, \gamma_n)^t$, donde $A = (\alpha_{ij})$ y $\mathbf{b} = (\beta_1, \ldots, \beta_n)^t$ se definen por medio de

$$\alpha_{ij} = \iint_{\mathcal{D}} \left[p(x, y)\frac{\partial \phi_i}{\partial x}(x, y)\frac{\partial \phi_j}{\partial x}(x, y) + q(x, y)\frac{\partial \phi_i}{\partial y}(x, y)\frac{\partial \phi_j}{\partial y}(x, y) \right. \tag{12.31}$$

$$\left. - r(x, y)\phi_i(x, y)\phi_j(x, y)\right] dx\, dy + \int_{\mathcal{S}_2} g_1(x, y)\phi_i(x, y)\phi_j(x, y)\, dS$$

para toda $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, m$, y

$$\beta_i = -\iint_{\mathcal{D}} f(x, y)\phi_i(x, y)\, dx\, dy + \int_{\mathcal{S}_2} g_2(x, y)\phi_i(x, y)\, dS - \sum_{k=n+1}^{m} \alpha_{ik}\gamma_k, \tag{12.32}$$

para toda $i = 1, \ldots, n$.

La elección de las funciones base es importante, porque una elección adecuada a menudo hace definida positiva y de banda a la matriz A. En el problema de segundo orden (12.26) suponemos que $\mathcal{D}$ es poligonal y que $\mathcal{S}$ es un conjunto contiguo de líneas rectas, de modo que $\mathcal{D} = D$.

Para iniciar el procedimiento dividimos la región D en un conjunto de triángulos $T_1, T_2, \ldots, T_M$ en el que el i-ésimo triángulo tiene tres vértices, o nodos, denotados con

$$V_j^{(i)} = (x_j^{(i)}, y_j^{(i)}), \quad \text{para } j = 1, 2, 3.$$

Con el fin de simplificar la notación escribimos $V_j^{(i)}$ simplemente como $V_j = (x_j, y_j)$ cuando trabajamos con el triángulo fijo T_i. Con cada vértice V_j asociamos un polinomio lineal

$$N_j^{(i)}(x, y) \equiv N_j(x, y) = a_j + b_jx + c_jy, \quad \text{donde} \quad N_j^{(i)}(x_k, y_k) = \begin{cases} 1, & \text{si } j = k, \\ 0, & \text{si } j \neq k. \end{cases}$$

Esto produce sistemas lineales de la forma

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} a_j \\ b_j \\ c_j \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

donde el elemento 1 ocurre en el j-ésimo renglón del vector de la derecha (en este caso, $j = 2$).

Sean $E_1, \ldots, E_n$ etiquetas de los nodos que se encuentran en $D \cup \mathcal{S}$ en forma de derecha a izquierda, de arriba abajo. Con cada nodo E_k, asociamos una función ϕ_k que es lineal en cada triángulo, que tiene el valor 1 en E_k, y que en el resto de los nodos es 0. Esta opción hace ϕ_k idéntica a $N_j^{(i)}$ en el triángulo T_i cuando el nodo E_k es el vértice denotado por $V_j^{(i)}$.

EJEMPLO 1 Supongamos que un problema de elemento finito contiene los triángulos T_1 y T_2 que aparecen en la figura 12.13. La función lineal $N_1^{(1)}(x, y)$ que asume el valor 1 en (1, 1) y 0 tanto en (0, 0) como en (−1, 2) satisface

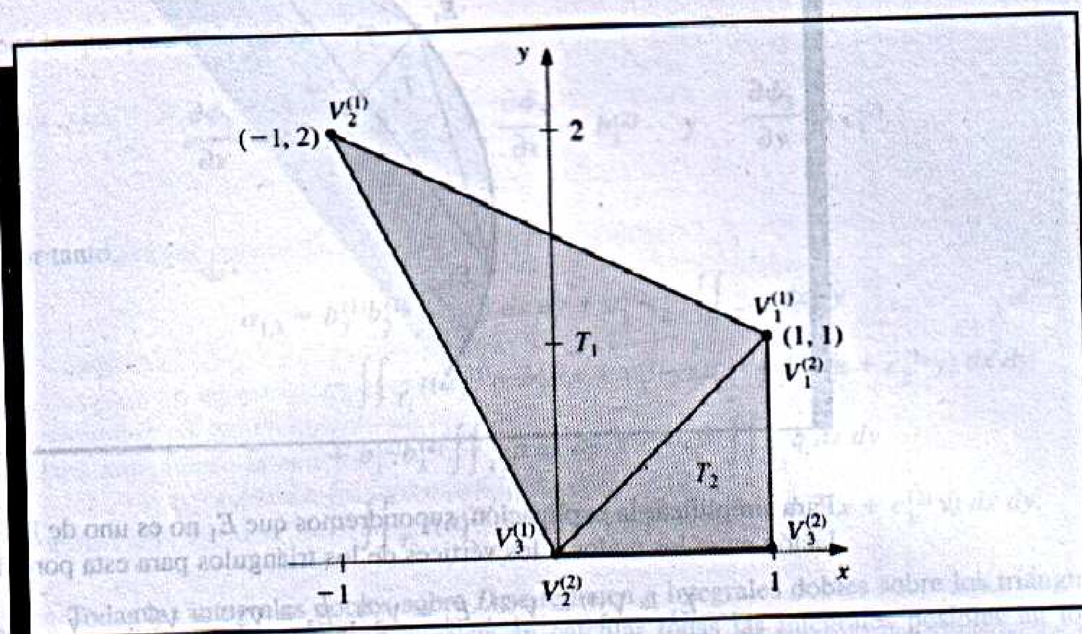
$$a_1^{(1)} + b_1^{(1)}(1) + c_1^{(1)}(1) = 1,$$

$$a_1^{(1)} + b_1^{(1)}(-1) + c_1^{(1)}(2) = 0,$$

y

$$a_1^{(1)} + b_1^{(1)}(0) + c_1^{(1)}(0) = 0.$$

Figura 12.13



Por tanto, $a_1^{(1)} = 0$, $b_1^{(1)} = \frac{2}{3}$, $c_1^{(1)} = \frac{1}{3}$, y

$$N_1^{(1)}(x, y) = \frac{2}{3}x + \frac{1}{3}y.$$

De igual forma, la función lineal $N_1^{(2)}(x, y)$ que asume el valor 1 en (1, 1) y 0 tanto en (0, 0) como en (1, 0) satisface

$$a_1^{(2)} + b_1^{(2)}(1) + c_1^{(2)}(1) = 1,$$

$$a_1^{(2)} + b_1^{(2)}(0) + c_1^{(2)}(0) = 0,$$

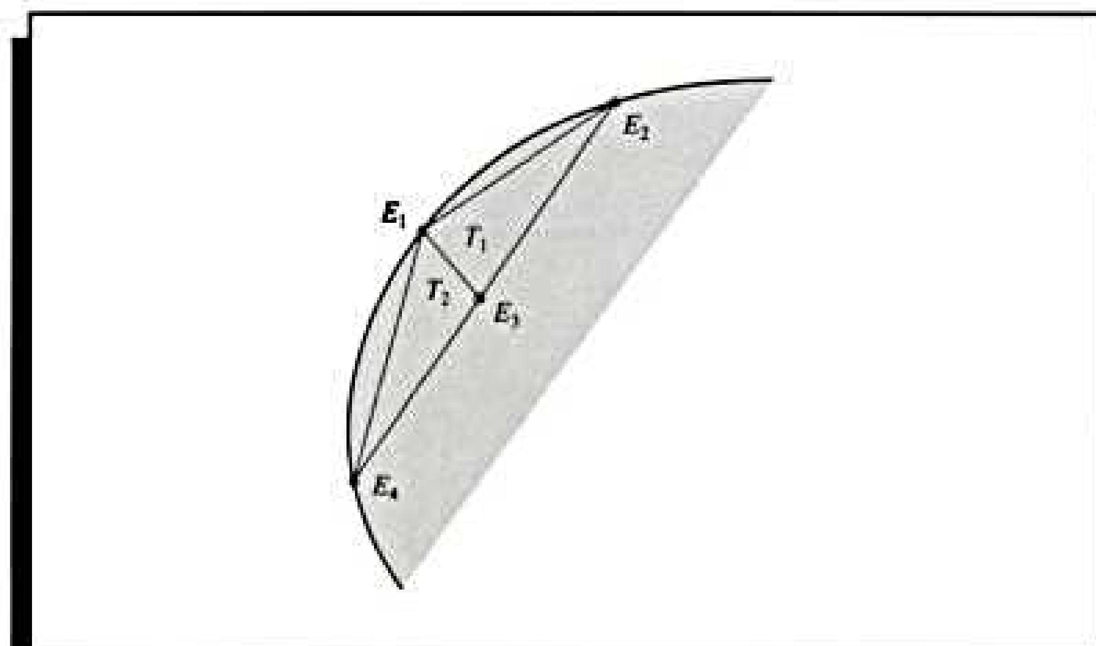
y

$$a_1^{(2)} + b_1^{(2)}(1) + c_1^{(2)}(0) = 0,$$

así que $a_1^{(2)} = 0$, $b_1^{(2)} = 0$ y $c_1^{(2)} = 1$. En consecuencia, $N_1^{(2)}(x, y) = y$. Nótese que en la frontera común de T_1 y T_2, $N_1^{(1)}(x, y) = N_1^{(2)}(x, y)$ ya que $y = x$. ■

Consideremos la Fig. 12.14, que es la parte superior izquierda de la región que se muestra en la Fig. 12.12. Generaremos los elementos de la matriz A que correspondan a los nodos que se muestran en esta figura.

Figura 12.14



Para simplificar la explicación, supondremos que E_1 no es uno de los nodos de $\mathcal{S}_2$. La relación entre los nodos y los vértices de los triángulos para esta porción es

$$E_1 = V_3^{(1)} = V_1^{(2)},\ E_4 = V_2^{(2)},\ E_3 = V_2^{(1)} = V_3^{(2)}, \quad \text{y} \quad E_2 = V_1^{(1)}.$$

Como ϕ_1 y ϕ_3 son cero en T_1 y T_2, los elementos $\alpha_{1,3} = \alpha_{3,1}$ se calculan mediante

$$\alpha_{1,3} = \iint_D \left[p \frac{\partial \phi_1}{\partial x} \frac{\partial \phi_3}{\partial x} + q \frac{\partial \phi_1}{\partial y} \frac{\partial \phi_3}{\partial y} - r\phi_1\phi_3 \right] dx\,dy$$

$$= \iint_{T_1} \left[p \frac{\partial \phi_1}{\partial x} \frac{\partial \phi_3}{\partial x} + q \frac{\partial \phi_1}{\partial y} \frac{\partial \phi_3}{\partial y} - r\phi_1\phi_3 \right] dx\,dy$$

$$+ \iint_{T_2} \left[p \frac{\partial \phi_1}{\partial x} \frac{\partial \phi_3}{\partial x} + q \frac{\partial \phi_1}{\partial y} \frac{\partial \phi_3}{\partial y} - r\phi_1\phi_3 \right] dx\,dy.$$

En el triángulo T_1,

$$\phi_1(x, y) = N_3^{(1)}(x, y) = a_3^{(1)} + b_3^{(1)}x + c_3^{(1)}y$$

y

$$\phi_3(x, y) = N_2^{(1)}(x, y) = a_2^{(1)} + b_2^{(1)}x + c_2^{(1)}y.$$

así, para toda (x, y),

$$\frac{\partial \phi_1}{\partial x} = b_3^{(1)}, \; \frac{\partial \phi_1}{\partial y} = c_3^{(1)}, \; \frac{\partial \phi_3}{\partial x} = b_2^{(1)}, \quad \text{y} \quad \frac{\partial \phi_3}{\partial y} = c_2^{(1)}.$$

De manera parecida, en T_2,

$$\phi_1(x, y) = N_1^{(2)}(x, y) = a_1^{(2)} + b_1^{(2)} x + c_1^{(2)} y$$

y

$$\phi_3(x, y) = N_3^{(2)}(x, y) = a_3^{(2)} + b_3^{(2)} x + c_3^{(2)} y,$$

por lo que para todo (x, y),

$$\frac{\partial \phi_1}{\partial x} = b_1^{(2)}, \; \frac{\partial \phi_1}{\partial y} = c_1^{(2)}, \; \frac{\partial \phi_3}{\partial x} = b_3^{(2)} \quad \text{y} \quad \frac{\partial \phi_3}{\partial y} = c_3^{(2)}.$$

Por tanto,

$$\begin{aligned}
\alpha_{1,3} = {} & b_3^{(1)} b_2^{(1)} \iint_{T_1} p \, dx \, dy + c_3^{(1)} c_2^{(1)} \iint_{T_1} q \, dx \, dy \\
& - \iint_{T_1} r (a_3^{(1)} + b_3^{(1)} x + c_3^{(1)} y)(a_2^{(1)} + b_2^{(1)} x + c_2^{(1)} y) \, dx \, dy \\
& + b_1^{(2)} b_3^{(2)} \iint_{T_2} p \, dx \, dy + c_1^{(2)} c_3^{(2)} \iint_{T_2} q \, dx \, dy \\
& - \iint_{T_2} r (a_1^{(2)} + b_1^{(2)} x + c_1^{(2)} y)(a_3^{(2)} + b_3^{(2)} x + c_3^{(2)} y) \, dx \, dy.
\end{aligned}$$

Todas las integrales dobles sobre D se reducen a integrales dobles sobre los triángulos. El procedimiento habitual consiste en calcular todas las integrales posibles en los triángulos y en acumularlas en el elemento correcto α_{ij} de A. De igual manera, las integrales dobles de la forma

$$\iint_D f(x, y) \phi_i(x, y) \, dx \, dy$$

se calculan en los triángulos y luego se acumulan en el elemento correcto β_i del vector **b**. Por ejemplo, para determinar β_1, necesitamos

$$\begin{aligned}
-\iint_D f(x, y) \phi_1(x, y) \, dx \, dy = {} & -\iint_{T_1} f(x, y) [a_3^{(1)} + b_3^{(1)} x + c_3^{(1)} y] \, dx \, dy \\
& - \iint_{T_2} f(x, y) [a_1^{(2)} + b_1^{(2)} x + c_1^{(2)} y] \, dx \, dy.
\end{aligned}$$

Parte de β_1 proviene de ϕ_1 restringida a T_1 y el resto proviene de ϕ_1 restringida a T_2 puesto que E_1 es un vértice tanto de T_1 como de T_2. Además, los nodos que se encuentran en $\mathcal{S}_2$ tienen incorporadas integrales de línea en sus elementos de A y **b**.

El algoritmo 12.5 aplica el método del elemento finito a la ecuación diferencial elíptica de segundo orden. Este algoritmo inicialmente asigna 0 a todos los valores de la ma-

triz A y del vector **b** y, una vez efectuadas todas las integraciones en los triángulos, agrega esos valores a los elementos correspondientes de A y **b**.

ALGORITMO 12.5

Método del elemento finito

Para aproximar la solución de la ecuación diferencial parcial

$$\frac{\partial}{\partial x}\left(p(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(q(x, y)\frac{\partial u}{\partial y}\right) + r(x, y)u = f(x, y), \quad (x, y) \in D$$

sujeta a las condiciones de frontera

$$u(x, y) = g(x, y), \quad (x, y) \in \mathcal{S}_1$$

y

$$p(x, y)\frac{\partial u}{\partial x}(x, y)\cos\theta_1 + q(x, y)\frac{\partial u}{\partial y}(x, y)\cos\theta_2 + g_1(x, y)u(x, y) = g_2(x, y),$$
$$(x, y) \in \mathcal{S}_2,$$

donde $\mathcal{S}_1 \cup \mathcal{S}_2$ es la frontera de D y θ_1 y θ_2 son los ángulos de dirección de la normal a la frontera:

Paso 0 Divida la región D en triángulos $T_1, \ldots, T_M$ tales que:
$T_1, \ldots, T_k$ sean los triángulos sin aristas en $\mathcal{S}_1$ y $\mathcal{S}_2$;
(*Nota: $K = 0$ implica que ningún triángulo es interior a D.*)
$T_{k+1}, \ldots, T_N$ sean los triángulos con, al menos, una arista en $\mathcal{S}_2$;
$T_{N+1}, \ldots, T_M$ sean los triángulos restantes.
(*Nota: $M = N$ implica que todos los triángulos tienen aristas en $\mathcal{S}_2$.*)
Etiquete los tres vértices del triángulo T_i como

$$\left(x_1^{(i)}, y_1^{(i)}\right), \left(x_2^{(i)}, y_2^{(i)}\right) \text{ y } \left(x_3^{(i)}, y_3^{(i)}\right).$$

Marque los nodos (vértices) $E_1, \ldots, E_m$ donde
$E_1, \ldots, E_m$ están en $D \cup \mathcal{S}_2$ y $E_{n+1}, \ldots, E_m$ están en $\mathcal{S}_1$.
(*Nota: $n = m$ implica que $\mathcal{S}_1$ no contiene nodos.*)

ENTRADA enteros K, N, M, n, m; vértices $\left(x_1^{(i)}, y_1^{(i)}\right), \left(x_2^{(i)}, y_2^{(i)}\right), \left(x_3^{(i)}, y_3^{(i)}\right)$ para cada $i = 1, \ldots, M$; nodos E_j para cada $j = 1, \ldots, m$.
(*Nota: Lo único que se necesita es un medio para establecer la correspondencia entre un vértice $\left(x_k^{(i)}, y_k^{(i)}\right)$ y un nodo $E_j = (x_j, y_j)$.*)

SALIDA constantes $\gamma_1, \ldots, \gamma_m$; $a_j^{(i)}, b_j(i), c_j(i)$ para cada $j = 1, 2, 3$ e $i = 1, \ldots, M$.

Paso 1 Para $l = n + 1, \ldots, m$, tome $\gamma_l = g(x_l, y_l)$. (*Nota: $E_l = (x_l, y_l)$.*)

Paso 2 Para $i = 1, \ldots, n$
tome $\beta_i = 0$;
para $j = 1, \ldots, n$, tome $\alpha_{i,j} = 0$.

Paso 3 Para $i = 1, \ldots, M$

tome $\Delta_i = \det \begin{vmatrix} 1 & x_1^{(i)} & y_1^{(i)} \\ 1 & x_2^{(i)} & y_2^{(i)} \\ 1 & x_3^{(i)} & y_3^{(i)} \end{vmatrix}$;

$$a_1^{(i)} = \frac{x_2^{(i)}y_3^{(i)} - y_2^{(i)}x_3^{(i)}}{\Delta_i}; \quad b_1^{(i)} = \frac{y_2^{(i)} - y_3^{(i)}}{\Delta_i}; \quad c_1^{(i)} = \frac{x_3^{(i)} - x_2^{(i)}}{\Delta_i};$$

$$a_2^{(i)} = \frac{x_3^{(i)}y_1^{(i)} - y_3^{(i)}x_1^{(i)}}{\Delta_i}; \quad b_2^{(i)} = \frac{y_3^{(i)} - y_1^{(i)}}{\Delta_i}; \quad c_2^{(i)} = \frac{x_1^{(i)} - x_3^{(i)}}{\Delta_i};$$

$$a_3^{(i)} = \frac{x_1^{(i)}y_2^{(i)} - y_1^{(i)}x_2^{(i)}}{\Delta_i}; \quad b_3^{(i)} = \frac{y_1^{(i)} - y_2^{(i)}}{\Delta_i}; \quad c_3^{(i)} = \frac{x_2^{(i)} - x_1^{(i)}}{\Delta_i};$$

para $j = 1, 2, 3$
defina $N_j^{(i)}(x, y) = a_j^{(i)} + b_j^{(i)}x + c_j^{(i)}y$.

Paso 4 Para $i = 1, \ldots, M$ (*Las integrales en los pasos 4 y 5 se pueden evaluar utilizando una integración numérica.*)
para $j = 1, 2, 3$
para $k = 1, \ldots, j$ (*Calcule todas las integrales dobles en los triángulos.*)

$$\text{tome } z_{j,k}^{(i)} = b_j^{(i)} b_k^{(i)} \int_{T_i}\!\!\int p(x, y)\, dx\, dy + c_j^{(i)} c_k^{(i)} \int_{T_i}\!\!\int q(x, y)\, dx\, dy$$
$$- \int_{T_i}\!\!\int r(x, y) N_j^{(i)}(x, y) N_k^{(i)}(x, y)\, dx\, dy;$$
$$\text{tome } H_j^{(i)} = - \int_{T_i}\!\!\int f(x, y) N_j^{(i)}(x, y)\, dx\, dy.$$

Paso 5 Para $i = K + 1, \ldots, N$ (*Calcule todas las integrales lineales.*)
para $j = 1, 2, 3$
para $k = 1, \ldots, j$

$$\text{tome } J_{j,k}^{(i)} = \int_{\mathcal{S}_2} g_1(x, y) N_j^{(i)}(x, y) N_k^{(i)}(x, y)\, dS;$$
$$\text{tome } I_j^{(i)} = \int_{\mathcal{S}_2} g_2(x, y) N_j^{(i)}(x, y)\, dS.$$

Paso 6 Para $i = 1, \ldots, M$, haga los pasos 7-12. (*Incorpore las integrales en cada triángulo al sistema lineal.*)

Paso 7 Para $k = 1, 2, 3$ haga los pasos 8-12.

Paso 8 Encuentre l tal que $E_l = \left(x_k^{(i)}, y_k^{(i)}\right)$.

Paso 9 Si $k > 1$, entonces para $j = 1, \ldots, k - 1$ haga los pasos 10, 11.

Paso 10 Encuentre t tal que $E_t = \left(x_j^{(i)}, y_j^{(i)}\right)$.

Paso 11 Si $l \leq n$, entonces
si $t \leq n$, entonces tome $\alpha_{lt} = \alpha_{lt} + z_{k,j}^{(i)}$;
$\alpha_{tl} = \alpha_{tl} + z_{k,j}^{(i)}$
si no, tome $\beta_l = \beta_l - \gamma_t z_{k,j}^{(i)}$
si no
si $t \leq n$, entonces tome $\beta_t = \beta_t - \gamma_l z_{k,j}^{(i)}$.

Paso 12 Si $l \leq n$, entonces tome $\alpha_{ll} = \alpha_{ll} + z_{k,j}^{(i)}$;
$\beta_l = \beta_l + H_k^{(i)}$.

Paso 13 Para $i = K + 1, \ldots, N$ haga los pasos 14-19. (*Incorpore las integrales de línea al sistema lineal.*)

Paso 14 Para $k = 1, 2, 3$ haga los pasos 15-19.

Paso 15 Encuentre l tal que $E_l = \left(x_k^{(i)}, y_k^{(i)}\right)$.

Paso 16 Si $k > 1$, entonces para $j = 1, \ldots, k - 1$ haga los pasos 17, 18.

Paso 17 Encuentre t tal que $E_t = \left(x_j^{(i)}, y_j^{(i)}\right)$.

Paso 18 Si $l \leq n$, entonces
si $t \leq n$, entonces tome $\alpha_{lt} = \alpha_{lt} + J_{k,j}^{(i)}$;
$\alpha_{tl} = \alpha_{tl} + J_{k,j}^{(i)}$
si no, tome $\beta_l = \beta_l - \gamma_t J_{k,j}^{(i)}$
si no
si $t \leq n$, entonces tome $\beta_t = \beta_t - \gamma_l J_{k,j}^{(i)}$.

Paso 19 Si $l \leq n$, entonces tome $\alpha_{ll} = \alpha_{ll} + J_{k,k}^{(i)}$;
$\beta_l = \beta_l - I_k^{(i)}$.

Paso 20 Resuelva el sistema lineal $A\mathbf{c} = \mathbf{b}$ donde $A = (\alpha_{l,t})$, $\mathbf{b} = (\beta_l)$ y $\mathbf{c} = (\gamma_t)$ para $1 \leq l \leq n$ y $1 \leq t \leq n$.

Paso 21 SALIDA $(\gamma_1, \ldots, \gamma_m)$.
(*Para cada* $k = 1, \ldots, m$ *sea* $\phi_k = N_j^{(i)}$ en T_i si $E_k = \left(x_j^{(i)}, y_j^{(i)}\right)$.
Entonces $\phi(x, y) = \sum_{k=1}^{m} \gamma_k \phi_k(x, y)$ aproxima $u(x, y)$ en $D \cup \mathcal{S}_1 \cup \mathcal{S}_2$.)

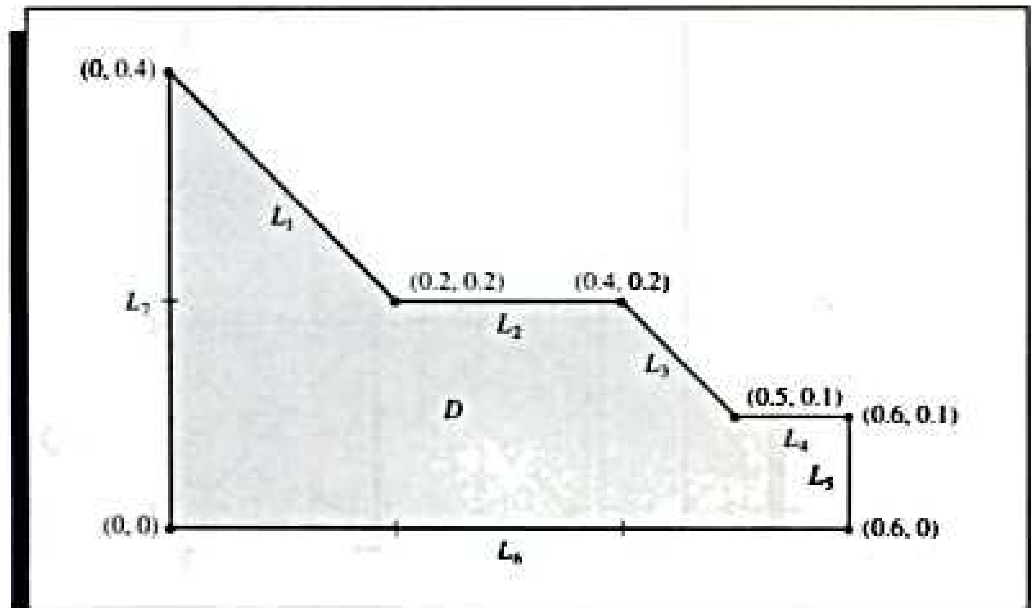
Paso 22 Para $i = 1, \ldots, M$
para $j = 1, 2, 3$ SALIDA $\left(a_j^{(i)}, b_j^{(i)}, c_j^{(i)}\right)$.

Paso 23 PARAR. (*Procedimiento terminado.*) ■

EJEMPLO 2 En una región bidimensional D la temperatura $u(x, y)$ satisface la ecuación de Laplace

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = 0 \text{ en } D.$$

Figura 12.15



Considere la región D de la Fig. 12.15 y suponga que se dan las siguientes condiciones de frontera:

$$u(x, y) = 4, \qquad \text{para } (x, y) \in L_6 \text{ y } (x, y) \in L_7;$$

$$\frac{\partial u}{\partial n}(x, y) = x, \qquad \text{para } (x, y) \in L_2 \text{ y } (x, y) \in L_4;$$

$$\frac{\partial u}{\partial n}(x, y) = y, \qquad \text{para } (x, y) \in L_5;$$

$$\frac{\partial u}{\partial n}(x, y) = \frac{x + y}{\sqrt{2}}, \qquad \text{para } (x, y) \in L_1 \text{ y } (x, y) \in L_3;$$

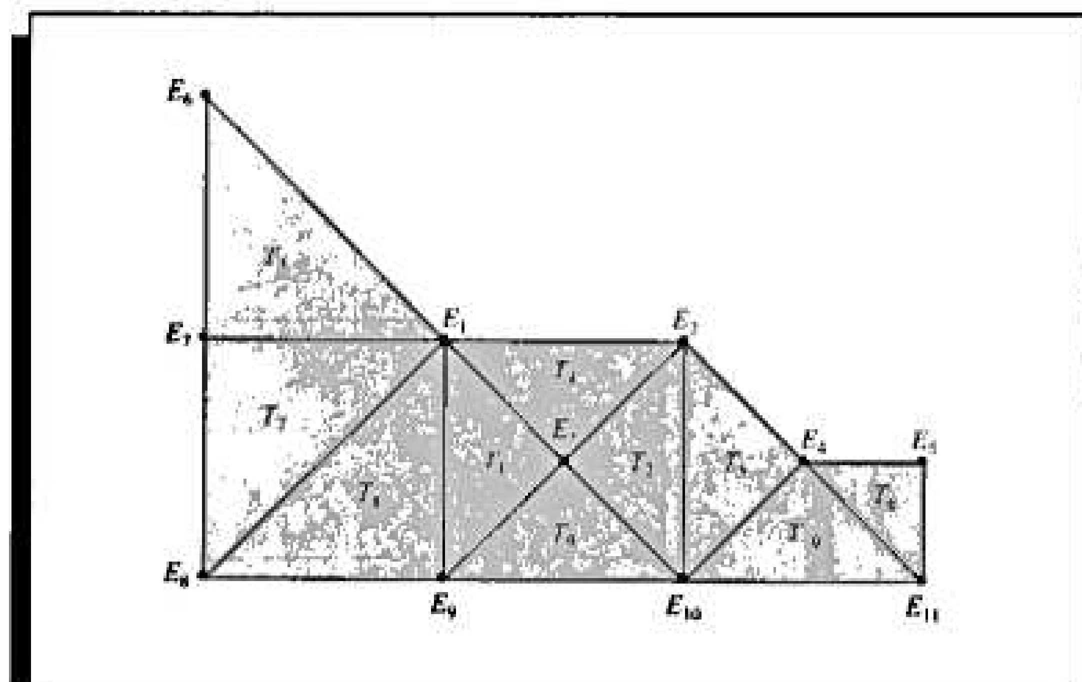
donde $\partial u/\partial n$ denota la derivada direccional en la dirección de la normal a la frontera de la región D en el punto (x, y).

Primero subdividimos D en triángulos etiquetados con las marcas que se sugieren en el paso 0 del algoritmo. En este ejemplo, $\mathcal{S}_1 = L_6 \cup L_7$ y $\mathcal{S}_2 = L_1 \cup L_2 \cup L_3 \cup L_4 \cup L_5$. Las marcas o etiquetas de los triángulos se muestran en la Fig. 12.16.

La condición de frontera $u(x, y) = 4$ en L_6 y L_7 implican que $\gamma_t = 4$ cuando $t = 6, 7, \ldots, 11$. Si queremos determinar los valores de γ_l para $l = 1, 2, \ldots, 5$, aplicamos los pasos restantes del algoritmo y generamos la matriz

$$A = \begin{bmatrix} 2.5 & 0 & -1 & 0 & 0 \\ 0 & 1.5 & -1 & -0.5 & 0 \\ -1 & -1 & 4 & 0 & 0 \\ 0 & -0.5 & 0 & 2.5 & -0.5 \\ 0 & 0 & 0 & -0.5 & 1 \end{bmatrix}$$

Figura 12.16



y el vector

$$\mathbf{b} = \begin{bmatrix} 6.066\bar{6} \\ 0.063\bar{3} \\ 8.0000 \\ 6.056\bar{6} \\ 2.031\bar{6} \end{bmatrix}.$$

La solución de la ecuación $A\mathbf{c} = \mathbf{b}$ es

$$\mathbf{c} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{bmatrix} = \begin{bmatrix} 4.0383 \\ 4.0782 \\ 4.0291 \\ 4.0496 \\ 4.0565 \end{bmatrix},$$

que da la siguiente aproximación a la solución de la ecuación de Laplace y las condiciones de frontera en los triángulos respectivos:

$$T_1: \quad \phi(x, y) = 4.0383(1 - 5x + 5y) + 4.0291(-2 + 10x) + 4(2 - 5x - 5y),$$

$$T_2: \quad \phi(x, y) = 4.0782(-2 + 5x + 5y) + 4.0291(4 - 10x) + 4(-1 + 5x - 5y),$$

$$T_3: \quad \phi(x, y) = 4(-1 + 5y) + 4(2 - 5x - 5y) + 4.0383(5x),$$

$$T_4: \quad \phi(x, y) = 4.0383(1 - 5x + 5y) + 4.0782(-2 + 5x + 5y) + 4.0291(2 - 10y),$$

$$T_5: \quad \phi(x, y) = 4.0782(2 - 5x + 5y) + 4.0496(-4 + 10x) + 4(3 - 5x - 5y),$$

$$T_6: \quad \phi(x, y) = 4.0496(6 - 10x) + 4.0565(-6 + 10x + 10y) + 4(1 - 10y),$$

$$T_7: \quad \phi(x, y) = 4(-5x + 5y) + 4.0383(5x) + 4(1 - 5y),$$
$$T_8: \quad \phi(x, y) = 4.0383(5y) + 4(1 - 5x) + 4(5x - 5y),$$
$$T_9: \quad \phi(x, y) = 4.0291(10y) + 4(2 - 5x - 5y) + 4(-1 + 5x - 5y),$$
$$T_{10}: \quad \phi(x, y) = 4.0496(10y) + 4(3 - 5x - 5y) + 4(-2 + 5x - 5y).$$

La solución real al problema con valor en frontera es $u(x, y) = xy + 4$. En la tabla 12.7 de la siguiente página se compara el valor de u con el de ϕ en E_i para toda $i = 1, \ldots, 5$. ■

Tabla 12.7

x	y	$\phi(x, y)$	$u(x, y)$	$\lvert \phi(x, y) - u(x, y) \rvert$
0.2	0.2	4.0383	4.04	0.0017
0.4	0.2	4.0782	4.08	0.0018
0.3	0.1	4.0291	4.03	0.0009
0.5	0.1	4.0496	4.05	0.0004
0.6	0.1	4.0565	4.06	0.0035

Por lo regular, el error de los problemas elípticos de segundo orden como el de (12.26) con coeficientes de funciones suaves es de $O(h^2)$, donde h es el diámetro máximo de los elementos triangulares. También se espera que las funciones base bilineales seccionadas en los elementos rectangulares den resultados de $O(h^2)$, donde h es la longitud de la diagonal máxima de esos elementos. Podemos usar otras clases de funciones base para obtener resultados de $O(h^4)$, pero su construcción es más compleja. Es difícil formular y aplicar los teoremas del error eficiente en los métodos del elemento finito, pues la exactitud de la aproximación depende de las propiedades de continuidad de la solución y de la regularidad de la frontera.

El método de elementos finitos también puede aplicarse a las ecuaciones diferenciales parabólicas e hiperbólicas, pero el procedimiento de minimización es más difícil. Una buena explicación de las ventajas y técnicas de su aplicación a varios problemas físicos se encuentra en un trabajo de [Fi]. Si el lector desea una exposición más amplia, le recomendamos consultar a [SF], [ZM] o [AB].

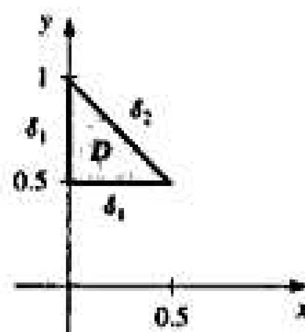
CONJUNTO DE EJERCICIOS 12.4

1. Use el algoritmo 12.5 para aproximar la solución de la siguiente ecuación diferencial parcial (véase la figura anexa):

$$\frac{\partial}{\partial x}\left(y^2 \frac{\partial u}{\partial x}(x, y)\right) + \frac{\partial}{\partial y}\left(y^2 \frac{\partial u}{\partial y}(x, y)\right) - yu(x, y) = -x, \quad (x, y) \in D,$$

$$u(x, 0.5) = 2x, \quad 0 \le x \le 0.5, \quad u(0, y) = 0, \quad 0.5 \le y \le 1,$$

$$y^2 \frac{\partial u}{\partial x}(x, y) \cos\theta 1 + y^2 \frac{\partial u}{\partial y}(x, y) \cos\theta_2 = \frac{\sqrt{2}}{2}(y - x) \quad \text{para } (x, y) \in \mathcal{S}_2.$$



Sea $M = 2$; T_1 tiene los vértices (0, 0.5), (0.25, 0.75), (0, 1); y T_2 tiene los vértices (0, 0.5), (0.5, 0.5) y (0.25, 0.75).

2. Repita el ejercicio 1 usando ahora los triángulos

T_1: (0, 0.75), (0, 1), (0.25, 0.75);
T_2: (0.25, 0.5), (0.25, 0.75), (0.5, 0.5);
T_3: (0, 0.5), (0, 0.75), (0.25, 0.75).
T_4: (0, 0.5), (0.25, 0.5), (0.25, 0.75).

3. Aproxime la solución de la ecuación diferencial parcial

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) - 12.5\pi^2 u(x, y) = -25\pi^2 \operatorname{sen}\frac{5\pi}{2}x \operatorname{sen}\frac{5\pi}{2}y, \quad 0 < x, y < 0.4,$$

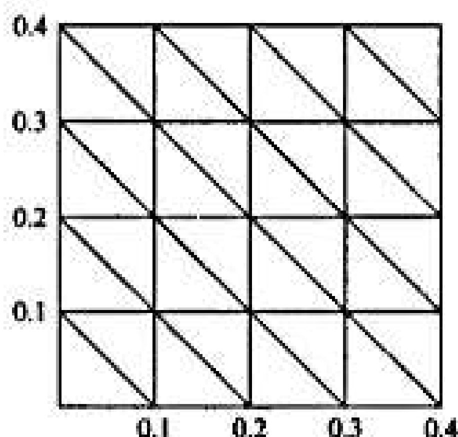
sujeta a la condición de frontera de Dirichlet

$$u(x, y) = 0,$$

usando el algoritmo de elementos finitos con los elementos que se incluyen en la figura anexa. Compare la solución aproximada con la solución real

$$u(x, y) = \operatorname{sen}\frac{5\pi}{2}x \operatorname{sen}\frac{5\pi}{2}y,$$

en los vértices interiores y en los puntos (0.125, 0.125), (0.125, 0.25), (0.25, 0.125) y (0.25, 0.25).



4. Repita el ejercicio 3 con $f(x, y) = -25\pi^2 \cos\frac{5\pi}{2}x \cos\frac{5\pi}{2}y$ usando la condición de frontera de Neumann

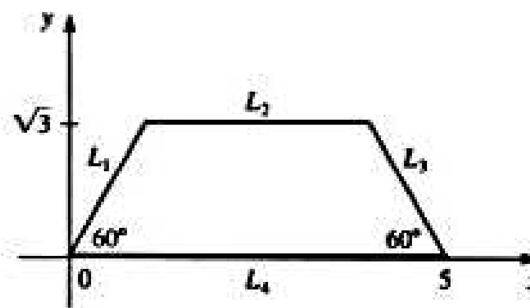
$$\frac{\partial u}{\partial n}(x, y) = 0.$$

La solución real de este problema es

$$u(x, y) = \cos\frac{5\pi}{2}x \cos\frac{5\pi}{2}y.$$

5. En una placa de plata en forma de trapecio (véase la figura anexa) se genera calor en todos los puntos con una rapidez de $q = 1.5$ cal/cm^3·s. La temperatura en estado estable $u(x, y)$ de la placa satisface la ecuación de Poisson

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = \frac{-q}{k},$$



donde k, la conductividad térmica, es 1.04 cal/cm·deg·s. Suponga que la temperatura se mantiene en 15°C en L_2, que se pierde calor en los bordes inclinados L_1 y L_3 conforme a la condición de frontera $\partial u/\partial n = 4$, y que no se pierde calor en L_4; es decir, $\partial u/\partial n = 0$. Use el algoritmo 12.5 para aproximar la temperatura de la placa en (1, 0), (4, 0) y $(\frac{5}{2}, \sqrt{3}/2)$.

12.5 Reseña de métodos y de software

En este capítulo explicamos los métodos con que se aproximan las soluciones de las ecuaciones diferenciales parciales. Limitamos nuestra atención a la ecuación de Poisson como ejemplo de una ecuación diferencial parcial elíptica, a la ecuación de calor o de difusión como ejemplo de una ecuación diferencial parcial parabólica, y a la ecuación de onda como ejemplo de una ecuación parcial hiperbólica. En estos tres ejemplos expusimos las aproximaciones de diferencias finitas.

La ecuación de Poisson en un rectángulo requiere la resolución de un sistema lineal grande y disperso para el cual recomendamos usar los métodos iterativos, como el SOR. Describimos cuatro métodos de diferencias finitas para la ecuación de calor. Los métodos de diferencias progresivas y de Richardson presentaron problemas de estabilidad, de ahí que hayamos explicado también los métodos de diferencias regresivas y de Crank-Nicolson. Aunque el sistema lineal tridiagonal debe resolverse en cada paso de tiempo con las técnicas implícitas, éstas son más estables que los métodos explícitos de diferencias pro-

gresivas y de Richardson. El método de diferencias finitas para la ecuación de onda es explícito y también puede presentar problemas de estabilidad para ciertas selecciones de discretizaciones en el tiempo y el espacio.

En la última sección del capítulo ofrecimos una introducción al método del elemento finito para una ecuación diferencial parcial elíptica autoadjunta en un dominio poligonal. Aunque nuestros métodos dan resultados satisfactorios en los problemas y ejemplos que se tratan en este libro, en las aplicaciones comerciales se requieren generalizaciones y modificaciones más poderosas de ellos.

Consideramos dos subrutinas de la biblioteca IMSL. La subrutina MOLCH sirve para resolver la ecuación diferencial parcial con las condiciones de frontera

$$\frac{\partial u}{\partial t} = F\left(x, t, u, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}\right),$$

con condiciones frontera

$$\alpha(x, t)u(x, t) + \beta(x, t)\frac{\partial u}{\partial x}(x, t) = \gamma(x, t).$$

El método se basa en la colocación de los puntos gaussianos en el eje x de cada valor de t, y utiliza como funciones base a los trazadores de Hermite.

La subrutina FPS2H se usa para resolver la ecuación de Poisson en un rectángulo. El método de resolución tiene como fundamento una elección de diferencias finitas de segundo o cuarto orden en una red uniforme.

La biblioteca NAG tiene varias subrutinas para las ecuaciones diferenciales parciales. La subrutina D03EAF se usa para la ecuación de Laplace en un dominio arbitrario del plano xy. La subrutina D03PAF sirve para resolver una sola ecuación diferencial parcial parabólica mediante el método de líneas.

Hay programas especializados, como NASTRAN, que contienen códigos para el método de elementos finitos, y que se utilizan comúnmente en aplicaciones de ingeniería. El paquete FISHPACK en la biblioteca Netlib sirve para resolver ecuaciones diferenciales parciales elípticas separables. Resulta difícil escribir códigos generales para las ecuaciones diferenciales parciales, por el problema que implica especificar los dominios que no son figuras geométricas comunes. Hoy en día se realiza mucha investigación en el área de la resolución de las ecuaciones diferenciales parciales.

Hemos descrito únicamente una pequeña muestra de los muchos métodos con que se aproximan las soluciones a los problemas que incluyen ecuaciones diferenciales parciales. Información más completa sobre el tema general se da en Lapidus y Pinder [LP], en Twizell [Tw], y en el libro reciente de Morton y Mayers [MM]. En Rice y Boisvert [RB] y en Bank [Ban] se encuentra información sobre los programas de cómputo.

Entre los libros dedicados al estudio de los métodos de diferencias finitas citamos los de Strikwerda [Strik], Thomas [Th] y Shashkov y Steinberg [ShS]. Strange y Fix [SF] y Zienkiewicz y Morgan [ZM] son buenas fuentes de información sobre el método de elementos finitos. Las ecuaciones dependientes del tiempo se estudian en Schiesser [Schi] y en Gustafsson, Kreiss y Oliger [GKO]; Birkhoff y Lynch [BL] y Roache [Ro] explican la solución a los problemas elípticos.

Los métodos multirredes utilizan las aproximaciones de red y los métodos iterativos para ofrecer aproximaciones en redes más finas. Entre las obras en que se explican esas técnicas figuran las de Briggs [Brigg], Mc Cormick [Mc] y Bramble [Bram].

Bibliografía

Las páginas del texto que se relacionan con cada referencia aparecen en cursivas.

[AHU] Aho, A. V., J. E. Hopcroft y J. D. Ullman, *The design and analysis of computer algorithms*, Addison-Wesley, Reading, MA, 1974, 470 pp. QA76.6A36 *546*

[AG] Allgower, E. y K. Georg, *Numerical continuation methods: an introduction*, Springer-Verlag, Nueva York, 1990, 388 pp. QA377.A56 *644*

[Am] Ames, W. F., *Numerical methods for partial differential equations* (tercera edición), Academic Press, Nueva York, 1992, 451 pp. QA374.A46 *724*

[An] Anderson, E., *et al.*, *LAPACK user's guide* (tercera edición), SIAM Publications, Philadelphia, PA, 1999, 407 pp. QA76.73.F25 L36 *44*

[AS] Argyros, I. K. y F. Szidarovszky, *The theory and applications of iteration methods*, CRC Press, Boca Raton, FL, 1993, 355 pp. QA297.8.A74 *644*

[AMR] Ascher, U. M., R. M. M. Mattheij y R. B. Russell, *Numerical solution of boundary value problems for ordinary differential equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988, 595 pp. QA379.A83 *690*

[Ax] Axelsson, O., *Iterative solution methods*, Cambridge University Press, Nueva York, 1994, 654 pp. QA297.8.A94 *482*

[AB] Axelsson, O. y V. A. Barker, *Finite element solution of boundary value problems: theory and computation*, Academic Press, Orlando, FL, 1984, 432 pp. QA379.A9 *739*

[Ba1] Bailey, N. T. J., *The mathematical approach to biology and medicine*, John Wiley & Sons, Nueva York, 1967, 269 pp. QH324.B28 *289*

[Ba2] Bailey, N. T. J., *The mathematical theory of epidemics*, Hafner, Nueva York, 1957, 194 pp. RA625.B3 *289*

[BSW] Bailey, P. B., L. F. Shampine y P. E. Waltman, *Nonlinear two-point boundary-value problems*, Academic Press, Nueva York, 1968, 171 pp. QA372.B27 *673, 690*

[Ban] Bank, R. E., *PLTMG, a software package for solving elliptic partial differential equation: Users' Guide 7.0*, SIAM Publications, Philadelphia, PA, 1994, 128 pp. QA377.B26 *742*

[Barr] Barrett, R., *et al.*, *Templates for the solution of linear systems: building blocks for iterative methods*, SIAM Publications, Philadelphia, PA, 1994, 112 pp. QA297.8.T45 *482*

[Bart] Bartle, R. G., *The elements of real analysis*, (segunda edición), John Wiley & Sons, Nueva York, 1976, 480 pp. QA300.B29 *105, 122*

[Bek] Bekker, M. G., *Introduction to terrain vehicle systems*, University of Michigan Press, Ann Arbor, MI, 1969, 846 pp. TL243.B39 *77, 620*

[Ber] Bernadelli, H., *Population waves*, Journal of the Burma Research Society **31** (1941), 1-18 DS527.B85 *381*

[BD] Birkhoff, G. y C. De Boor, *Error bounds for spline interpolation*, Journal of Mathematics and Mechanics **13** (1964), 827-836, QA1.J975 *152*

[BL] Birkhoff, G. y R. E. Lynch, *Numerical solution of elliptic problems*, SIAM Publications, Philadelphia, PA, 1984, 319 pp. QA377.B672 *742*

[BiR] Birkhoff, G. y G. Rota, *Ordinary differential equations* (cuarta edición), John Wiley & Sons, Nueva York, 1989, 399 pp. QA372.B58 *252, 254, 313*

[BP] Botha, J. F. y G. F. Pinder, *Fundamental concepts in the numerical solution of differential equations*, Wiley-Interscience, Nueva York, 1983, 202 pp. QA374.B74 *343*

[Brac] Bracewell, R., *The Fourier transform and its application* (tercera edición), McGraw-Hill, Nueva York, 2000, 616 pp. QA403.5.B7 *546*

[Bram] Bramble, J. H., *Multigrid methods*, John Wiley & Sons, Nueva York, 1993, 161 pp. QA377.B73 *742*

[Bre] Brent, R., *Algorithms for minimization without derivatives*, Prentice-Hall, Englewood Cliffs, NJ, 1973, 195 pp. QA402.5.B74 *102, 103*

[Brigg] Briggs, W. L., *A multigrid tutorial*, SIAM Publications, Philadelphia, PA, 1987, 88 pp. QA377.B75 *742*

[BH] Briggss, W. L. y V. E. Henson, *The DFT: an owner's manual for the discrete Fourier transform*, SIAM Publications, Philadelphia, PA, 1995, 434 pp. QA403.5.B75 *549*

[Brigh] Brigham, E. O., *The fast Fourier transform*, Prentice-Hall, Englewood Cliffs, NJ, 1974, 252 pp. QA403.B74 *538*

[Brow,K] Brown, K. M., *A quadratically convergent Newton-like method based upon Gaussian elimination*, SIAM Journal on Numerical Analysis **6**, No. 4 (1969), 560-569, QA297.A1S2 *626*

[Brow,W] Brown, W. S., *A simple but realistic model of floating point computation*, ACM transactions of Mathematical Software 7 (1981), 445-480, QA76.A8 *42, 43*

[Broy] Broyden, C. G., *A class of methods for solving nonlinear simultaneous equations*, Mathematics of Computation **19** (1965), 577-593, QA1.M4144 *621*

[BS1] Bulirsch R. y J. Stoer, *Numerical treatment of ordinary differential equations by extrapolation methods*, Numerische Mathematik **8** (1966), 1-13, QA241.N9 *312*

[BS2] Bulirsch, R. y J. Stoer, *Fehlerabschätzungen und extrapolation mit rationalen Funktionen bei Verfahren von Richardson-typus*, Numerische Mathematik **6** (1964), 413-427, QA241.N9 *312*

[BS3] Bulirsch, R. y J. Stoer, *Asymptotic upper and lower bounds for results of extrapolation methods*, Numerische Mathematik **8** (1966), 93-104, QA241.N9 *312*

[BuR] Bunch, J. R. y D. J. Rose (eds.), *Sparse matrix computations* (Proceedings of a conference held at Argonne National Laboratories, septiembre 9-11, 1975), Academic Press, Nueva York, 1976, 453 pp. QA188.S9 *414*

[BFR] Burden, R. L., J. D. Faires y A. C. Reynolds, *Numerical Analysis* (segunda edición), Prindle, Weber & Schmidt, Boston, MA, 1981, 598 pp. QA297.B84 *95*

[Bur] Burrage, K., 1995, *Parallel and sequential methods for ordinary differential equations*, Oxford University Press, Nueva York, 446 pp. QA372.B883 *343*

[But] Butcher, J. C., *The non-existence of ten-stage eighth-order explicit Runge-Kutta methods*, BIT **25** (1985), 521-542, QA76.N62 *279*

[CF] Chaitin-Chatelin, F. y Fraysse, V., *Lectures on finite precision computations*, SIAM Publications, Philadelphia, PA, 1996, 235 pp. QA297.C417 *46*

[CGGG] Char, B. W., K. O. Geddes, W. M. Gentlemen, G. H. Gonnet, *The design of Maple: A compact, portable, and powerful computer algebra system*, Computer Algebra. Lecture Notes in Computer Science No. 162 (J. A. Van Hulzen, ed.), Springer-Verlag, Berlin, 1983, 101-115 pp. QA155.7 E4 E85 *46*

[CCR] Chiarella, C., W. Charlton y A. W. Roberts, *Optimum chute profiles in gravity flow of granular materials: a discrete segment solution method*, Transactions of the ASME, Journal of Engineering for Industry Series B **97** (1975), 10-13, TJ1.A712 *619*

[Ch] Cheney, E. W., *Introduction to approximation theory*, McGraw-Hill, Nueva York, 1966, 259 pp. QA221.C47 *549*

[CC] Clenshaw, C. W. y C. W. Curtis, *A method for numerical integration on an automatic computer*, Numerische Mathematik **2** (1960), 197-205, QA241.N9 *248*

[CW] Cody, W. J. y W. Waite, *Software manual for the elementary functions*, Prentice-Hall, Englewood Cliffs, NJ, 1980, 269 pp. QA331.C635 *46*

[CV] Coleman, T. F. y C. Van Loan, *Handbook for matrix computations*, SIAM Publications, Philadelphia, PA, 1988, 264 pp. QA188.C65 *44, 416*

[CT] Cooley, J. W. y J. W. Tukey, *An algorithm for the machine calculation of complex Fourier series*, Mathematics of Computation **19**, No. 90 (1965), 297-301, QA1.M4144 *538*

[Co] Cowell, W. (ed.), *Sources and development of mathematical software*, Prentice-Hall, Englewood Cliffs, NJ, 1984, 404 pp. QA76.95.S68 *43*

[DaB] Dahlquist, G. y Å. Björck (Translated by N. Anderson), *Numerical methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974, 573 pp. QA297.D3313 *86*

[Da] Davis, P. J., *Interpolation and approximation*, Dover, Nueva York, 1975, 393 pp. QA221.D33 *165, 549*

[DR] Davis, P. J. y P. Rabinowitz, *Methods of numerical integration* (segunda edición), Academic Press, Nueva York, 1984, 612 pp. QA299.3.D28 *248*

[Deb] De Boor, C., *A practical guide to splines*, Springer-Verlag, Nueva York, 1978, 392 pp. QA1.A647 vol. 27 *152, 164, 165*

[DebS] De Boor, C. y B. Swartz, *Collocation at Gaussian points*, SIAM Journal on Numerical Analysis **10**, No. 4 (1973), 582-606, QA297.A1S2 *687*

[DM] Dennis, J. E., Jr. y J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Review **19**, No. 1 (1977), 46-89, QA1.S2 *622*

[DenS] Dennis, J. E., Jr. y R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983, 378 pp. QA402.5.D44 *644*

[Di] Dierckx, P., *Curve and surface fitting with splines*, Oxford University Press, Nueva York, 1993, 285 pp. QA297.6.D54 *165*

[DBMS] Dongarra, J. J., J. R. Bunch, C. B. Moler y G. W. Stewart, *LINPACK users guide*, SIAM Publications, Philadelphia, PA, 1979, 367 pp. QA214.L56 *44*

[DRW] Dongarra, J. J., T. Rowan y R. Wade, *Software distributions using Xnetlib*, ACM Transactions on Mathematical Software **21**, No. 1 (1995), 79-88 QA76.6.A8 *44*

[DW] Dongarra, J. y D. W. Walker, *Software libraries for linear algebra computation on high performance computers*, SIAM Review **37**, No. 2 (1995), 151-180 QA1.S2 *46*

[Do] Dormand, J. R., *Numerical methods for differential equations: a computational approach*, CRC Press, Boca Raton, FL, 1996, 368 pp. QA372.D67 *343*

[DoB] Dorn, G. L. y A. B. Burdick, *On the recombinational structure of complementation relationships in the m-dy complex of the Drosophila melanogaster*, Genetics **47** (1962), 503-518, QH431.G43 *418*

[E] Engels, H., *Numerical quadrature and cubature*, Academic Press, Nueva Yok, 1980, 441 pp. QA299.3.E5 *248*

[Fe] Fehlberg, E., *Klassische Runge-Kutta Formeln vierter und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme*, Computing **6** (1970), 61-71, QA76.C777 *284*

[Fi] Fix, G., *A survey of numerical methods for selected problems in continuum mechanics*, Proceedings of a Conference on Numerical Methods of Ocean Circulation, National Academy of Sciences (1975), 268-283, Q11.N26 *739*

[FM] Forsythe, G. E. y C. B. Moler, *Computer solution of linear algebraic systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967, 148 pp. QA297.F57 *416, 457*

[Fr] Francis, J. G. F., *The QR transformation*, Computer Journal **4** (1961-1962), Part I, 265-271; Part II, 332-345, QA76.C57 *589*

[Fu] Fulks, W., *Advanced calculus* (tercera edición), John Wiley & Sons, Nueva York, 1978, 731 pp. QA303.F954 *10, 272*

[Gar] Garbow, B. S., *et al.*, *Matrix eigensystem routines: EISPACK guide extension*, Springer-Verlag, Nueva York, 1977, 343 pp. QA193.M38 *44*

[Ge1] Gear, C. W., *Numerical initial-value problems in ordinary differential equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971, 253 pp. QA372.G4 *322, 325, 336, 343*

[Ge2] Gear, C. W., *Numerical solution of ordinary differential equations: Is there anything left to do?*, SIAM Review **23**, No. 1 (1981), 10-24, QA1.S2 *340*

[GL] George, A. y J. W. Liu, *Computer solution of large sparse positive definite systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981, 324 pp. QA188.G46 *416*

[Go] Goldberg, D., *What every scientist should know about floating-point arithmetic*, ACM Computing Surveys **23**, No. 1 (1991), 5-48, QA76.5.A1 *46*

[GO] Golub, G. H. y Ortega, J. M., *Scientific computing: an introduction with parallel computing*, Academic Press, Boston, MA, 1993, 442 pp. QA76.58.G64 *46, 343*

[GV] Golub, G. H. y C. F. Van Loan, *Matrix computations* (segunda edición), Johns Hopkins University Press, Baltimore, MD, 1989, 642 pp. QA188.G65 *400, 416*

[Gr] Gragg, W. B., *On extrapolation algorithms for ordinary initial-value problems*, SIAM Journal on Numerical Analysis **2** (1965), 384-403, QA297.A1S2 *307, 312*

[GKO] Gustafsson, B., H. Kreiss y J. Oliger, *Time dependent problems and difference methods*, John Wiley & Sons, Nueva York, 1995, 642 pp. QA374.G974 *742*

[Hac] Hackbusch, W., *Iterative solution of large sparse systems of equations*, Springer-Verlag, Nueva York, 1994, 429 pp. QA1.A647 vol. 95 *482*

[HY] Hageman, L. A. y D. M. Young, *Applied iterative methods*, Academic Press, Nueva York, 1981, 386 pp. QA297.8.H34 *482*

[HNW1] Hairer, E., S. P. Nörsett y G. Wanner, *Solving ordinary differential equations. Vol. 1: Nonstiff equations* (segunda edición revisada), Springer-Verlag, Berlín, 1993, 519 pp. QA372.H16 *322, 343*

[HNW2] Hairer, E., S. P. Nörsett y G. Wanner, *Solving ordinary differential equations. Vol. 2: Stiff and differential-algebraic problems* (segunda edición revisada), Springer, Berlín, 1996, 614 pp. QA372.H16 *343*

[Ham] Hamming, R. W., *Numerical methods for scientists and engineers* (segunda edición), McGraw-Hill, Nueva York, 1973, 721 pp. QA297.H28 *546*

[He1] Henrici, P., *Discrete variable methods in ordinary differential equations*, John Wiley & Sons, Nueva York, 1962, 407 pp. QA372.H48 *343*

[He2] Henrici, P., *Elements of numerical analysis*, John Wiley & Sons, Nueva York, 1964, 328 pp. QA297.H54 *90, 330*

[HS] Hestenes, M. R. y E. Steifel, *Conjugate gradient methods in optimization*, Journal of Research of the National Bureau of Standards **49** (1952), 409-436, Q1.N34 *465*

[Hild] Hildebrand, F. B., *Introduction to numerical analysis* (segunda edición), McGraw-Hill, Nueva York, 1974, 669 pp. QA297.H54 *122, 124, 129*

[Hill] Hill, F. S., Jr., *Computer graphics: using openGL* (segunda edición), Prentice-Hall, Englewood Cliffs, NJ, 2001, 922 pp. T385.H549 *162*

[Ho] Householder, A. S., *The numerical treatment of a single nonlinear equation*, McGraw-Hill, Nueva York, 1970, 216 pp. QA218.H68 *102, 103*

[IK] Issacson, E. y H. B. Keller, *Analysis of numerical methods*, John Wiley & Sons, Nueva York, 1966, 541 pp. QA297.I8 *90, 192, 193, 328, 331, 436, 563, 565, 652, 670, 707, 711, 713, 724*

[JT] Jenkins, M. A. y J. F. Traub, *A three-stage algorithm for real polynomials using quadratic iteration*, SIAM Journal on Numerical Analysis **7**, No. 4 (1970), 545-556, QA297.A1S2 *102*

[Joh] Johnston, R. L., *Numerical methods: a software approach*, John Wiley & Sons, Nueva York, 1982, 276 pp. QA297.J64 *211*

[Joy] Joyce, D. C., *Survey of extrapolation processes in numerical analysis*, SIAM Review **13**, No. 4 (1971), 435-490, QA1.S2 *178*

[Keller,H] Keller, H. B., *Numerical methods for two-point boundary-value problems*, Blaisdell, Waltham, MA, 1968, 184 pp. QA372.K42 *659, 664, 668, 689, 690*

[Keller,J] Keller, J. B., *Probability of a shutout in racquetball*, SIAM Review **26**, No. 2 (1984), 267-268, QA1.S2 *77*

[Kelley] Kelley, C. T., *Iterative methods for linear and nonlinear equations*, SIAM Publications, Philadelphia, PA, 1995, 165 pp. QA297.8.K45 *477, 482*

[Ko] Köckler, N., *Numerical methods and scientific computing: using software libraries for problem solving*, Oxford University Press, Nueva York, 1994, 328 pp. TA345.K653 *46*

[Lam] Lambert, J. D., *The initial value problem for ordinary differential equations. The state of art in numerical analysis* (D. Jacobs, ed.), Academic Press, Nueva York, 1977, 451-501 pp. QA297.C646 *340*

[LP] Lapidus, L. y G. F. Pinder, *Numerical solution of partial differential equations in science and engineering*, John Wiley & Sons, Nueva York, 1982, 677 pp. Q172.L36 *742*

[Lar] Larson, H. J., *Introduction to probability theory and statistical inference* (tercera edición), John Wiley & Sons, Nueva York, 1982, 637 pp. QA273.L352 *486*

[Lau] Laufer, H. B., *Discrete mathematics and applied modern algebra*, PWS-Kent Publishing, Boston, MA, 1984, 583 pp. QA161.L38 *546*

[LH] Lawson, C. L. y Hanson, R. J., *Solving least squares problems*, SIAM Publications, Philadelphia, PA, 1995, 337 pp. QA275.L38 *549*

[LR] Lucas, T. R. y G. W. Reddien, Jr., *Some collocation methods for nonlinear boundary value problems*, SIAM Journal on Numerical Analysis **9**, No. 2 (1972), 341-356, QA297.A1S2 *687*

[Lu] Luenberger, David G., *Linear and nonlinear programming* (segunda edición), Addison-Wesley, Reading, MA, 1984, 245 pp. T57.7L8 *471*

[Ma] Mano, M. M., *Computer system architecture*, Prentice-Hall, Englewood Cliffs, NJ, 1982, 531 pp. QA76.9A73 M36 *22*

[Mc] McCormick, S. F., *Multigrid methods*, SIAM Publications, Philadelphia, PA, 1987, 282 pp. QA374.M84 *742*

[Mi] Mitchell, A. R., *Computation methods in partial differential equations*, John Wiley & Sons, Nueva York, 1969, 255 pp. QA374.M68 *724*

[Mo] Moler, C. B., *Demonstration of a matrix laboratory. Lecture notes in Mathematics* (J. P. Hennart, ed.), Springer-Verlag, Berlín, 1982, 84-98 *45*

[MC] Moré J. J. y M. Y. Cosnard, *Numerical solution of nonlinear equations*, ACM Transactions on Mathematical Software **5**, No. 1 (1979), 64-85, QA76.6.A8 *626*

[MM] Morton, K. W. y D. F. Mayers, *Numerical solution of partial differential equations: an introduction*, Cambridge University Press, Nueva York, 1994, 227 pp. QA377.M69 *742*

[Mu] Müller, D. E., *A method for solving algebraic equations using an automatic computer*, Mathematical Tables and Other Aids to Computation **10** (1956), 208-215, QA47.M29 *95*

[ND] Noble, B. y J. W. Daniel, *Applied linear algebra* (tercera edición), Prentice-Hall, Englewood Cliffs, NJ, 1988, 521 pp. QA184.N6 *370, 384, 551*

[Or1] Ortega, J. M., *Introduction to parallel and vector solution of linear systems*, Plenum Press, Nueva York, 1988, 305 pp. QA218.O78 *45*

[Or2] Ortega, J. M., *Numerical analysis; a second course*, Academic Press, Nueva York, 1972, 201 pp. QA297.O78 *424, 434, 445, 449, 461, 557, 601, 604*

[OP] Ortega, J. M. y W. G. Poole, Jr., *An introduction to numerical methods for differential equations*, Pitman Publishing, Marshfield, MA, 1981, 329 pp. QA371.O65 *343*

[OR] Ortega, J. M. y W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, Nueva York, 1970, 572 pp. QA297.8.O77 *601, 634, 636, 644*

[Os] Ostrowski, A. M., *Solution of equations and systems of equations* (segunda edición), Academic Press, Nueva York, 1966, 338 pp. QA3.P8 vol. 9 *103*

[Par] Parlett, B., *The symmetric eigenvalue problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980, 348 pp. QA188.P37 *599*

[Pat] Patterson, T. N. L., *The optimum addition of points to quadrature formulae*, Mathematics of Computation **22**, No. 104 (1968), 847-856, QA1.M4144 *248*

[PF] Phillips, C. y T. L. Freeman, *Parallel numerical algorithms*, Prentice-Hall, Nueva York, 1992, 315 pp. QA76.9.A43 F74 *46*

[Ph] Phillips, J., *The NAG Library: a beginner's guide*, Clarendon Press, Oxford, 1986, 245 pp. QA297.P35 *45*

[PDUK] Piessens, R., E. de Doncker-Kapenga, C. W. Überhuber y D. K. Kahaner, *QUADPACK: a subroutine package for automatic integration*, Springer-Verlag, Nueva York, 1983, 301 pp. QA299.3.Q36 *247*

[Pi] Pissanetzky, S., *Sparse matrix technology*, Academic Press, Nueva York, 1984, 321 pp. QA188.P57 *416*

[Po] Powell, M. J. D., *Approximation theory and methods*, Cambridge University Press, Cambridge, 1981, 339 pp. QA221.P65 *137, 138, 165, 527, 549*

[Pr] Pryce, J. D., *Numerical solution of Sturm-Liouville problems*, Oxford University Press, Nueva York, 1993, 322 pp. QA379.P79 *690*

[RR] Ralston, A. y P. Rabinowitz, *A first course in numerical analysis* (segunda edición), McGraw-Hill, Nueva York, 1978, 556 pp. QA297.R3 *209, 523, 527, 634*

[Ra] Rashevsky, N., *Looking at history through mathematics*, Massachusetts Institute of Technology Press, Cambridge, MA, 1968, 199 pp. D16.25.R3 *268*

[RB] Rice, J. R. y R. F. Boisvert, *Solving elliptic problems using ELLPACK*, Springer-Verlag, Nueva York, 1985, 497 pp. QA377.R53 *742*

[RG] Richardson, L. F. y J. A. Gaunt, *The deferred approach to the limit*, Philosophical Transactions of the Royal Society of London **226A** (1927), 299-361, Q41.L82 *178*

[Ro] Roache, P. J., *Elliptic marching methods and domain decomposition*, CRC Press, Boca Raton, FL, 1995, 190 pp. QA377.R63 *742*

[RS] Roberts, S. y J. Shipman, *Two-point boundary value problems: shooting methods*, Elsevier, Nueva York, 1972, 269 pp. QA372.R76 *690*

[RW] Rose, D. J. y R. A. Willoughby (eds.), *Sparse matrices and their applications* (Proceedings of a conference held at IBM Research, Nueva York, septiembre 9-10, 1971, 215 pp.), Plenum Press, Nueva York, 1972, QA263.S94 *414*

[Ru] Russell, R. D., *A comparison of collocation and finite differences for two-point boundary value problems*, SIAM Journal on Numerical Analysis **14**, No. 1 (1977), 19-39, QA297.A152 *687*

[Sa1] Saad, Y., *Numerical methods for large eigenvalue problems*, Halsted Press, Nueva York, 1992, 346 pp. QA188.S18 *587*

[Sa2] Saad, Y., *Iterative methods for sparse linear systems*, PWS-Kent Publishing, Boston, MA, 1996, 447 pp. QA188.S17 *481, 482*

[SaS] Saff, E. B. y A. D. Snider, *Fundamentals of complex analysis for mathematics, science, and engineering* (segunda edición), Prentice-Hall, Englewood Cliffs, NJ, 1993, 468 pp. QA300.S18 *91*

[SP] Sagar, V. y D. J. Payne, *Incremental collapse of thick-walled circular cylinders under steady axial tension and torsion loads and cyclic transient heating*, Journal of the Mechanics and Physics of Solids **21**, No. 1 (1975), 39-54, TA350.J68 *717*

[SD] Sale, P. F. y R. Dybdahl, *Determinants of community structure for coral-reef fishes in experimental habitat*, Ecology **56** (1975), 1343-1355, QH540.E3 *496*

[Sche] Schendel, U., *Introduction to numerical methods for parallel computers* (Translated by B.W. Conolly), Halsted Press, Nueva York, 1984, 151 pp. QA297.S3813 *46*

[Schi] Schiesser, W. E., *Computational mathematics in engineering and applied science; ODE's, DAE's, and PDE's*, CRC Press, Boca Raton, FL, 1994, 587 pp. TA347.D45 S34 *742*

[Scho] Schoenberg, I. J., *Contributions to the problem of approximation of equidistant data by analytic functions*, Quarterly of Applied Mathematics **4** (1976), Part A, 45-99; Part B, 112-141, QA1.A26 *165*

[Schr1] Schroeder, L. A., *Energy budget of the larvae of the moth Pachysphinx modesta*, Oikos **24** (1973), 278-281, QH540.O35 *497*

[Schr2] Schroeder, L. A., *Thermal tolerances and acclimation of two species of hydras*, Limnology and Oceanography **26**, No. 4 (1981), 690-696, GC1.L5 *620*

[Schul] Schultz, M. H., *Spline analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973, 156 pp. QA211.S33 *152, 165, 673, 681*, 686, 687

[Schum] Schumaker, L. L., *Spline functions: basic theory*, Wiley-Interscience, Nueva York, 1981, 553 pp. QA224.S33 *165*

[Se] Searle, S. R., *Matrix algebra for the biological sciences*, John Wiley & Sons, Nueva York, 1966, 296 pp. QH324.S439 *381*

[SH] Secrist, D. A. y R. W. Hornbeck, *An analysis of heat transfer and fade in disk brakes*, Transactions of the ASME, Journal of Engineering for Industry Series B **98** No. 2 (1976), 385-390, TJ1.A712 *205*

[Sh] Shampine, L. F., *Numerical solution of ordinary differential equations*, Chapman & Hall, Nueva York, 1994, 484 pp. QA372.S417 *343*

[SGe] Shampine, L. F. y C. W. Gear, *A user's view of solving stiff ordinary differential equations*, SIAM Review **21**, No. 1 (1979), 1-17, QA1.S2 *340*

[ShS] Shashkov, M. y S. Steinberg, *Conservative finite-difference methods on general grids*, CRC Press, Boca Raton, FL, 1996, 359 pp. QA431.S484 *742*

[SW] Simon, B. y R. M. Wilson, *Supercalculators on the PC*, Notices of the American Mathematical Society **35**, No. 7 (1988), 978-1001, QA1.N6 *46*

[Si] Singh, V. P., *Investigations of attenuation and internal friction of rocks by ultrasonics*, International Journal of Rock Mechanics and Mining Sciences (1976), 69-72, TA706.I45 *496*

[SJ] Sloan, I. H. y S. Joe, *Lattice methods for multiple integration*, Oxford University Press, New York, 1994, 239 pp. QA311.S56 *248*

[Sm, B] Smith, B. T., *et al.*, *Matrix eigensystem routines: EISPACK guide* (segunda edición), Springer-Verlag, Nueva York, 1976, 551 pp. QA193.M37 *44*

[Sm.G] Smith, G. D., *Numerical solution of partial differential equations*, Oxford University Press, Nueva York, 1965, 179 pp. QA377.S59 *724*

[So] Sorenson, Danny C., *Parallel numerical algorithms* (David E. Keyes, Ahmed Sameh y V. Venkatakrishan, eds.), Kluwer Academic Publishers Dordrecht, 1997, 119 pp. QA76.9.A43 P35 *599*

[Stet] Stetter, H. J., *Analysis of discretization for ordinary differential equations. From tracts in natural philosophy*, Springer-Verlag, Nueva York, 1973, 388 pp. QA372.S84 *312*

[Stew1] Stewart, G. W., *Afternotes on numerical analysis*, SIAM Publications, Philadelphia, PA, 1996, 200 pp. QA297.S785 *402, 416*

[Stew2] Stewart, G. W., *Introduction to matrix computations*, Academic Press, Nueva York, 1973, 441 pp. QA188.S7 *402, 599*

[SF] Strang, W. G. y G. J. Fix, *An analysis of the finite element method*, Prentice-Hall, Englewood Cliffs, NJ, 1973, 306 pp. TA335.S77 *687, 739, 742*

[Stri] Strikwerda, J. C., *Finite difference schemes and partial differential equations*, Brooks/Cole Publishing, Pacific Grove, CA, 1989, 386 pp. QA374.S88 *742*

[Stro] Stroud, A. H., *Approximate calculation of multiple integrals*, Prentice-Hall, Englewood Cliffs, NJ, 1971, 431 pp. QA311.S85 *248*

[StS] Stroud, A. H. y D. Secrest, *Gaussian quadrature formulas*, Prentice-Hall, Englewood Cliffs, NJ, 1966, 374 pp. QA299.4.G4 S7 *224, 248*

[Sz] Szüsz, P., *Math bite*, Mathematics Magazine 68, No. 2, 1995, 97, QA1.N28 *430*

[Th] Thomas, J. W., *Numerical partial differential equations*, Springer-Verlag, Nueva York, 1995, 445 pp. QA377.T495 *742*

[Tr] Traub, J. F., *Iterative methods for the solution of equations*, Prentice-Hall, Englewood Cliffs, NJ, 1964, 310 pp. QA297.T7 *103*

[Tw] Twizell, E. H., *Computational methods for partial differential equations*, Ellis Horwood Ltd., Chichester, West Sussex, England, 1984, 276 pp. QA377.T95 *742*

[Van] Van Loan, C. F., *Computational frameworks for the fast Fourier transform*, SIAM Publications, Philadelphia, PA, 1992, 273 pp. QA403.5.V35 *549*

[Var] Varga, R. S., *Matrix iterative analysis* (segunda edición), Springer, Nueva York, 2000, 358 pp. QA263.V3 *482, 696, 700, 701*

[We] Wendroff, B., *Theoretical numerical analysis*, Academic Press, Nueva York, 1966, 239 pp. QA297.W43 *400, 403*

[Wil1] Wilkinson, J. H., *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963, 161 pp. QA76.5.W53 *462*

[Wil2] Wilkinson, J. H., *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965, 662 pp. QA218.W5 *462, 570, 578, 594, 599*

[WR] Wilkinson, J. H. y C. Reinsch (eds.), *Handbook for automatic computation, Vol.2: Linear algebra*, Springer-Verlag, Nueva York, 1971, 439 pp. QA251.W67 *43, 590, 594, 595, 599*

[Win] Winograd, S., *On computing the discrete Fourier transform*, Mathematics of Computation **32** (1978), 175-199, QA1.M4144 *546*

[Y] Young, D. M., *Iterative solution of large linear systems*, Academic Press, Nueva York, 1971, 570 pp. QA195.Y68 *445, 482*

[YG] Young, D. M. and R. T. Gregory, *A survey of numerical mathematics. Vol. 1*, Addison-Wesley, Reading, MA, 1972, 533 pp. QA297.Y63 *102*

[ZM] Zienkiewics, O. C. y K. Morgan, *Finite elements and approximation*, John Wiley & Sons, Nueva York, 1983, 328 pp. QA297.5.Z53 *739, 742*

Respuestas a ejercicios seleccionados

Conjunto de ejercicios 1.1

1. Para cada parte, $f \in C[a, b]$ en el intervalo dado. Puesto que $f(a)$ y $f(b)$ tienen signo contrario, según el teorema del valor intermedio existe un número c con $f(c) = 0$.
3. Para cada parte, $f \in C[a, b]$, f' existe en (a, b) y $f(a) = f(b) = 0$. Según el teorema de Rolle, existe un número c en (a, b) con $f'(c) = 0$. Para el inciso (d) podemos usar $[a, b] = [-1, 0]$ o $[a, b] = [0, 2]$.
5. Para $x < 0$, $f(x) < 2x + k < 0$, a condición de que $x < -\frac{1}{2}k$. De modo parecido, para $x > 0$, $f(x) > 2x + k > 0$, a condición de que $x > -\frac{1}{2}k$. Conforme al teorema 1.13, existe un número c con $f(c) = 0$. Si $f(c) = 0$ y $f(c') = 0$ para alguna $c' \neq c$, entonces, según el teorema 1.7, existe un número p entre c y c' con $f'(p) = 0$. Pero $f'(x) = 3x^2 + 2 > 0$ para toda x.
7. a. $P_2(x) = 0$

   b. $R_2(0.5) = 0.125$; error real $= 0.125$

   c. $P_2(x) = 1 + 3(x - 1) + 3(x - 1)^2$

   d. $R_2(0.5) = -0.125$; error real $= -0.125$
9. Dado que
$$P_2(x) = 1 + x \quad \text{y} \quad R_2(x) = \frac{-2e^{\xi}(\text{sen}\,\xi + \cos\,\xi)}{6}x^3$$
para alguna ξ entre x y 0, tendremos lo siguiente:

   a. $P_2(0.5) = 1.5$ y $|f(0.5) - P_2(0.5)| \leq 0.0532$;

   b. $|f(x) - P_2(x)| \leq 1.252$;

   c. $\int_0^1 f(x)dx \approx 1.5$;

   d. $\left|\int_0^1 f(x)\,dx - \int_0^1 P_2(x)\,dx\right| \leq \int_0^1 |R_2(x)|\,dx \leq 0.313$, y el error real es 0.122.
11. $P_3(x) = (x - 1)^2 - \frac{1}{2}(x - 1)^3$

   a. $P_3(0.5) = 0.312500$, $f(0.5) = 0.346574$. Una cota de error es $0.291\overline{6}$, y el error real es 0.034074.

   b. $|f(x) - P_3(x)| \leq 0.291\overline{6}$ en $[0.5, 1.5]$

   c. $\int_{0.5}^{1.5} P_3(x)\,dx = 0.08\overline{3}$, $\int_{0.5}^{1.5}(x - 1)\ln x\,dx = 0.088020$

   d. Una cota de error es $0.058\overline{3}$, y el error real es 4.687×10^{-3}.
13. $P_4(x) = x + x^3$

   a. $|f(x) - P_4(x)| \leq 0.012405$

   b. $\int_0^{0.4} P_4(x)\,dx = 0.0864$, $\int_0^{0.4} xe^{x^2}\,dx = 0.086755$

c. 8.27×10^{-4}

d. $P_4'(0.2) = 1.12, f'(0.2) = 1.124076$. El error real es 4.076×10^{-3}.

15. Puesto que $42° = 7\pi/30$ radianes, usamos $x_0 = \pi/4$. Entonces,

$$\left|R_n\left(\frac{7\pi}{30}\right)\right| \le \frac{\left(\frac{\pi}{4} - \frac{7\pi}{30}\right)^{n+1}}{(n+1)!} < \frac{(0.053)^{n+1}}{(n+1)!}.$$

Para $\left|R_n(\frac{7\pi}{30})\right| < 10^{-6}$, es suficiente tomar $n = 3$. Con 7 dígitos, $\cos 42° = 0.7431448$ y $P_3(42°) = P_3(\frac{7\pi}{30}) = 0.7431446$, así que el error real será 2×10^{-7}.

17. a. $P_3(x) = \ln(3) + \frac{2}{3}(x-1) + \frac{1}{9}(x-1)^2 - \frac{10}{81}(x-1)^3$

b. $\text{máx}_{0 \le x \le 1} |f(x) - P_3(x)| = |f(0) - P_3(0)| = 0.02663366$

c. $\tilde{P}_3(x) = \ln(2) + \frac{1}{2}x^2$

d. $\text{máx}_{0 \le x \le 1} |f(x) - \tilde{P}_3(x)| = |f(1) - \tilde{P}_3(1)| = 0.09453489$

e. $P_3(0)$ aproxima $f(0)$ mejor que $\tilde{P}_3(1)$ aproxima $f(1)$.

19. $P_n(x) = \sum_{k=0}^{n} \frac{1}{k!} x^k, n \ge 7$

21. Una cota de error máximo es 0.0026.

23. Puesto que $R_2(1) = \frac{1}{6}e^{\xi}$ para alguna ξ en $(0, 1)$, tenemos $|E - R_2(1)| = \frac{1}{6}|1 - e^{\xi}| \le \frac{1}{6}(e-1)$.

25. a. Sea x_0 cualquier número en $[a, b]$. Con $\epsilon > 0$, sea $\delta = \epsilon/L$. Si $|x - x_0| < \delta$ y $a \le x \le b$, entonces $|f(x) - f(x_0)| \le L|x - x_0| < \epsilon$.

b. Al aplicar el teorema del valor medio, tenemos

$$|f(x_2) - f(x_1)| = |f'(\xi)||x_2 - x_1|,$$

para alguna ξ entre x_1 y x_2, así que

$$|f(x_2) - f(x_1)| \le L|x_2 - x_1|.$$

c. Un ejemplo es $f(x) = x^{1/3}$ en $[0, 1]$.

27. a. Puesto que f es continuo en p y $f(p) \ne 0$, existe una $\delta > 0$ con

$$|f(x) - f(p)| < \frac{|f(p)|}{2},$$

para $|x - p| < \delta$ y $a < x < b$. Restringimos δ de modo que $[p - \delta, p + \delta]$ sea un subconjunto de $[a, b]$. Así, para $x \in [p - \delta, p + \delta]$, tenemos $x \in [a, b]$. Por tanto,

$$-\frac{|f(p)|}{2} < f(x) - f(p) < \frac{|f(p)|}{2}$$

y

$$f(p) - \frac{|f(p)|}{2} < f(x) < f(p) + \frac{|f(p)|}{2}.$$

Si $f(p) > 0$, entonces

$$f(p) - \frac{|f(p)|}{2} = \frac{f(p)}{2} > 0, \quad \text{para} \quad f(x) > f(p) - \frac{|f(p)|}{2} > 0.$$

Si $f(p) < 0$, entonces $|f(p)| = f(p)$, y

$$f(x) < f(p) + \frac{|f(p)|}{2} = f(p) - \frac{f(p)}{2} = \frac{f(p)}{2} < 0.$$

En cualquier caso, $f(x) \ne 0$, para $x \in [p - \delta, p + \delta]$.

b. Puesto que f es continua en p y $f(p) = 0$, existe $\delta > 0$ con

$$|f(x) - f(p)| < k, \quad \text{para} \quad |x - p| < \delta \quad \text{y} \quad a < x < b.$$

Restringimos δ de modo que $[p - \delta, p + \delta]$ sea un subconjunto de $[a, b]$. Así, para $x \in [p - \delta, p + \delta]$, tenemos

$$|f(x)| = |f(x) - f(p)| < k.$$

Conjunto de ejercicios 1.2

1.	Error absoluto	Error relativo
a.	0.001264	4.025×10^{-4}
b.	7.346×10^{-6}	2.338×10^{-6}
c.	2.818×10^{-4}	1.037×10^{-4}
d.	2.136×10^{-4}	1.510×10^{-4}
e.	2.647×10^{1}	1.202×10^{-3}
f.	1.454×10^{1}	1.050×10^{-2}
g.	420	1.042×10^{-2}
h.	3.343×10^{3}	9.213×10^{-3}

3. Los intervalos más grandes son **a.** (149.85, 150.15) **b.** (899.1, 900.9) **c.** (1498.5, 1501.5) **d.** (89.91, 90.09)

5.	Aproximación	Error absoluto	Error relativo
a.	134	0.079	5.90×10^{-4}
b.	133	0.499	3.77×10^{-3}
c.	2.00	0.327	0.195
d.	1.67	0.003	1.79×10^{-3}
e.	1.80	0.154	0.0786
f.	−15.1	0.0546	3.60×10^{-3}
g.	0.286	2.86×10^{-4}	10^{-3}
h.	0.00	0.0215	1.00

7.	Aproximación	Error absoluto	Error relativo
a.	133	0.921	6.88×10^{-3}
b.	132	0.501	3.78×10^{-3}
c.	1.00	0.673	0.402
d.	1.67	0.003	1.79×10^{-3}
e.	3.55	0.60	0.817
f.	−15.2	0.0454	0.00299
g.	0.284	0.00171	0.00600
h.	0	0.02150	1

9.	Aproximación	Error absoluto	Error relativo
a.	3.14557613	3.983×10^{-3}	1.268×10^{-3}
b.	3.14162103	2.838×10^{-5}	9.032×10^{-6}

11. a. $\lim_{x \to 0} \dfrac{x \cos x - \operatorname{sen} x}{x - \operatorname{sen} x} = \lim_{x \to 0} \dfrac{-x \operatorname{sen} x}{1 - \cos x} = \lim_{x \to 0} \dfrac{-\operatorname{sen} x - x \cos x}{\operatorname{sen} x} = \lim_{x \to 0} \dfrac{-2 \cos x + x \operatorname{sen} x}{\cos x} = -2$

b. −1.941

c. $\dfrac{x\left(1-\frac{1}{2}x^2\right)-\left(x-\frac{1}{6}x^3\right)}{x-\left(x-\frac{1}{6}x^3\right)}=-2$

d. El error relativo del inciso (b) es 0.029. El error relativo del inciso (c) es 0.00050.

13.

	x_1	Error absoluto	Error relativo	x_2	Error absoluto	Error relativo
a.	92.26	0.01542	1.672×10^{-4}	0.005419	6.273×10^{-7}	1.157×10^{-4}
b.	0.005421	1.264×10^{-6}	2.333×10^{-4}	−92.26	4.580×10^{-3}	4.965×10^{-5}
c.	10.98	6.875×10^{-3}	6.257×10^{-4}	0.001149	7.566×10^{-8}	6.584×10^{-5}
d.	−0.001149	7.566×10^{-8}	6.584×10^{-5}	−10.98	6.875×10^{-3}	6.257×10^{-4}

15. Los números de computadora equivalen a

a. 3224

b. −3224

c. 1.32421875

d. 1.3242187500000002220446049250313080847263336181640625

17. b. La primera fórmula da −0.00658 y la segunda fórmula da −0.0100. El valor verdadero de tres dígitos es −0.0116.

19. Las soluciones aproximadas de los sistemas son **a.** $x = 2.451$, $y = -1.635$ **b.** $x = 507.7$, $y = 82.00$

21. a. En la forma anidada tenemos

$$f(x) = (((1.01e^x - 4.62)e^x - 3.11)e^x + 12.2)e^x - 1.99.$$

b. −6.79. **c.** −7.07

23. a. $n = 77$ **b.** $n = 35$

25. a. $m = 17$

b. $\binom{m}{k} = \dfrac{m!}{k!(m-k)!} = \dfrac{m(m-1)\cdots(m-k-1)(m-k)!}{k!(m-k)!} = \left(\dfrac{m}{k}\right)\left(\dfrac{m-1}{k-1}\right)\cdots\left(\dfrac{m-k-1}{1}\right)$

c. $m = 181707$

d. 2 597 000; error real 1960; error relativo 7.541×10^{-4}

27. a. 124.03 **b.** 124.03 **c.** −124.03 **d.** −124.03 **e.** 0.0065 **f.** 0.0065 **g.** −0.0065 **h.** −0.0065

Conjunto de ejercicios 1.3

1. a. Las sumas aproximadas son 1.53 y 1.54, respectivamente. El valor real es 1.549. En el primer método se presenta un error significativo de redondeo.

3. a. 2000 términos **b.** 20 000 000 000 de términos

5. 3 términos

7. Las rapideces de convergencia son

a. $O(h^2)$ **b.** $O(h)$ **c.** $O(h^2)$ **d.** $O(h)$

13. a. Si $|\alpha_n - \alpha|/(1/n^p) \leq K$, entonces $|\alpha_n - \alpha| \leq K(1/n^p) \leq K(1/n^q)$ porque $0 < q < p$. Por tanto, $|\alpha_n - \alpha|/(1/n^p) \leq K$ y $\{\alpha_n\}_{n=1}^{\infty} \to \alpha$ con una rapidez de convergencia $O(1/n^p)$.

b.

n	$1/n$	$1/n^2$	$1/n^3$	$1/n^5$
5	0.2	0.04	0.008	0.0016
10	0.1	0.01	0.001	0.0001
50	0.02	0.0004	8×10^{-6}	1.6×10^{-7}
100	0.01	10^{-4}	10^{-6}	10^{-8}

La mayor rapidez de convergencia es $O(1/n^4)$.

15. Suponga que para una $|x|$ suficientemente pequeña, tenemos constantes positivas k_1 y k_2 independientes de x, para la que

$$|F_1(x) - L_1| \leq K_1|x|^{\alpha} \quad \text{y} \quad |F_2(x) - L_2| \leq K_2|x|^{\beta}$$

Sea $c = \text{máx}(|c_1|, |c_2|, 1)$, $K = \text{máx}(K_1, K_2)$ y $\delta = \text{máx}(\alpha, \beta)$.

a. Tenemos

$$\begin{aligned} |F(x) - c_1L_1 - c_2L_2| &= |c_1(F_1(x) - L_1) + c_2(F_2(x) - L_2)| \\ &\leq |c_1|K_1|x|^{\alpha} + |c_2|K_2|x|^{\beta} \\ &\leq cK[|x|^{\alpha} + |x|^{\beta}] \\ &\leq cK|x|^{\gamma}[1 + |x|^{\delta-\gamma}] \\ &\leq \tilde{K}|x|^{\gamma}, \end{aligned}$$

para una $|x|$ suficientemente pequeña y alguna constante $\tilde{K}$. Así, $F(x) = c_1L_1 + c_2L_2 + O(x^{\gamma})$.

b. Tenemos

$$\begin{aligned} |G(x) - L_1 - L_2| &= |F_1(c_1x) + F_2(c_2x) - L_1 - L_2| \\ &\leq K_1|c_1x|^{\alpha} + K_2|c_2x|^{\beta} \\ &\leq Kc^{\delta}[|x|^{\alpha} + |x|^{\beta}] \\ &\leq Kc^{\delta}|x|^{\gamma}[1 + |x|^{\delta-\gamma}] \\ &\leq \tilde{K}|x|^{\gamma}, \end{aligned}$$

para una $|x|$ suficientemente pequeña y alguna constante $\tilde{K}$. Así, $G(x) = L_1 + L_2 + O(x^{\gamma})$.

17. a. 354224848179261915075

b. $0.3542248538 \times 10^{21}$

c. El resultado del inciso (a) se calcula usando la aritmética de enteros exactos; el resultado del inciso (b) se calcula usando la aritmética con redondeo a 10 dígitos.

d. El resultado del inciso (a) requiere cruzar un ciclo 98 veces.

e. El resultado es el mismo que el del inciso (a).

Conjunto de ejercicios 2.1

1. $p_3 = 0.625$

3. El método de bisección da: **a.** $p_7 = 0.5859$ **b.** $p_8 = 3.002$ **c.** $p_7 = 3.419$

5. El método de bisección da $p_9 = 4.4932$.

7. El método de bisección da:

a. $p_{17} = 0.641182$

b. $p_{17} = 0.257530$

c. Para el intervalo $[-3, -2]$ tenemos $p_{17} = -2.191307$ y para el intervalo $[-1, 0]$ tenemos $p_{17} = -0.798164$.

d. Para el intervalo $[0.2, 0.3]$ tenemos $p_{14} = 0.297528$ y para el intervalo $[1.2, 1.3]$ tenemos $p_{14} = 1.256622$.

9. a. 2 **b.** -2 **c.** -1 **d.** 1

11. La tercera raíz de 25 es aproximadamente $p_{14} = 2.92401$, usando $[2, 3]$.

13. Una cota es $n \geq 14$ y $p_{14} = 1.32477$.

15. Puesto que $\lim_{n\to\infty}(p_n - p_{n-1}) = \lim_{n\to\infty} 1/n = 0$, la diferencia de términos es cero. Sin embargo, p_n es el n-ésimo término de la serie armónica divergente y, por tanto, $\lim_{n\to\infty} p_n = \infty$.

17. La profundidad del agua es 0.838 pies.

Conjunto de ejercicios 2.2

1. Para el valor de x en cuestión tenemos

a. $x = (3 + x - 2x^2)^{1/4} \Leftrightarrow x^4 = 3 + x - 2x^2 \Leftrightarrow f(x) = 0$

b. $x = \left(\dfrac{x + 3 - x^4}{2}\right)^{1/2} \Leftrightarrow 2x^2 = x + 3 - x^4 \Leftrightarrow f(x) = 0$

c. $x = \left(\dfrac{x + 3}{x^2 + 2}\right)^{1/2} \Leftrightarrow x^2(x^2 + 2) = x + 3 \Leftrightarrow f(x) = 0$

d. $x = \dfrac{3x^4 + 2x^2 + 3}{4x^3 + 4x - 1} \Leftrightarrow 4x^4 + 4x^2 - x = 3x^4 + 2x^2 + 3 \Leftrightarrow f(x) = 0$

3. El orden de la rapidez descendente de la convergencia es (b), (d), (a). La sucesión en (c) no converge.

5. Con $g(x) = (3x^2 + 3)^{1/4}$ y $p_0 = 1$, $p_6 = 1.94332$ tiene una exactitud de 0.01.

7. Puesto que $g'(x) = \frac{1}{4}\cos\frac{x}{2}$, g es continua y existe g' en $[0, 2\pi]$. Más aún, $g'(x) = 0$ solamente cuando $x = \pi$, así que $g(0) = g(2\pi) = \pi \le g(x) \le g(\pi) = \pi + \frac{1}{2}$ y $|g'(x)| \le \frac{1}{4}$ para $0 \le x \le 2\pi$. Según el teorema 2.2, existe un punto fijo p en $[0, 2\pi]$. Con $k = \frac{1}{4}$ y $p_0 = \pi$, tenemos $p_1 = \pi + \frac{1}{2}$. Conforme al corolario 2.4,

$$|p_n - p| \le \frac{k^n}{1-k}\,|p_1 - p_0| = \frac{2}{3}\left(\frac{1}{4}\right)^n.$$

Para que la cota sea menor que 0.1 necesitamos $n \ge 4$. Pero $p_3 = 3.626996$ tiene una exactitud de 0.01.

9. Para $p_0 = 1.0$ y $g(x) = 0.5(x + \frac{3}{x})$, tenemos $\sqrt{3} \approx p_4 = 1.73205$.

11. a. Con $[0, 1]$ y $p_0 = 0$, tenemos $p_9 = 0.257531$.

b. Con $[2.5, 3.0]$ y $p_0 = 2.5$, tenemos $p_{17} = 2.690650$.

c. Con $[0.25, 1]$ y $p_0 = 0.25$, tenemos $p_{14} = 0.909999$.

d. Con $[0.3, 0.7]$ y $p_0 = 0.3$, tenemos $p_{39} = 0.469625$.

e. Con $[0.3, 0.6]$ y $p_0 = 0.3$, tenemos $p_{48} = 0.448059$.

f. Con $[0, 1]$ y $p_0 = 0$, tenemos $p_6 = 0.704812$.

13. Para $g(x) = (2x^2 - 10\cos x)/(3x)$ tenemos lo siguiente:

$$p_0 = 3 \Rightarrow p_8 = 3.16193; \quad p_0 = -3 \Rightarrow p_8 = -3.16193.$$

Para $g(x) = \arccos(-0.1x^2)$, tenemos lo siguiente:

$$p_0 = 1 \Rightarrow p_{11} = 1.96882; \quad p_0 = -1 \Rightarrow p_{11} = -1.96882.$$

15. Con $g(x) = \frac{1}{\pi}\operatorname{arcsen}(-\frac{x}{2}) + 2$, tenemos $p_5 = 1.683855$.

17. Uno de los muchos ejemplos es $g(x) = \sqrt{2x - 1}$ en $[\frac{1}{2}, 1]$.

21. Reemplace la segunda oración de la demostración por: "Puesto que g satisface una condición de Lipschitz en $[a, b]$ con una constante de Lipschitz $L < 1$, para cada n tenemos

$$|p_n - p| = |g(p_{n-1}) - g(p)| \le L|p_{n-1} - p|."$$

El resto de la demostración es igual, sólo que k se sustituye con L.

23. Con $g(t) = 501.0625 - 201.0625e^{-0.4t}$ y $p_0 = 5.0$, $p_3 = 6.0028$ se encuentra dentro de 0.01 s del tiempo real.

Conjunto de ejercicios 2.3

1. $p_2 = 2.60714$

3. a. 2.45454 **b.** 2.44444 **c.** El inciso (b) es mejor.

5. a. Para $p_0 = 2$, tenemos $p_5 = 2.69065$.

b. Para $p_0 = -3$, tenemos $p_3 = -2.87939$.

c. Para $p_0 = 0$, tenemos $p_4 = 0.73909$.

d. Para $p_0 = 0$, tenemos $p_3 = 0.96434$.

7. Al usar los extremos de los intervalos como p_0 y p_1, tenemos:

i. a. $p_{11} = 2.69065$ **b.** $p_7 = -2.87939$ **c.** $p_6 = 0.73909$ **d.** $p_5 = 0.96433$

ii. a. $p_{16} = 2.69060$ **b.** $p_6 = -2.87938$ **c.** $p_7 = 0.73908$ **d.** $p_6 = 0.96433$

9. Para $p_0 = 1$, tenemos $p_5 = 0.589755$. El punto tiene las coordenadas (0.589755, 0.347811).

11. La ecuación de la tangente es

$$y - f(p_{n-1}) = f'(p_{n-1})(x - p_{n-1}).$$

Para completar este problema, utilice $y = 0$ y resuelva para $x = p_n$.

13. a. Para $p_0 = -1$ y $p_1 = 0$, tenemos $p_{17} = -0.04065850$ y para $p_0 = 0$ y $p_1 = 1$, tenemos $p_9 = 0.9623984$.

b. Para $p_0 = -1$ y $p_1 = 0$, tenemos $p_5 = -0.04065929$ y para $p_0 = 0$ y $p_1 = 1$, tenemos $p_{12} = -0.04065929$.

c. Para $p_0 = -0.5$, tenemos $p_5 = -0.04065929$ y para $p_0 = 0.5$, tenemos $p_{21} = 0.9623989$.

15. Esta fórmula incluye la sustracción de los números casi iguales en el numerador y en el denominador, si p_{n-1} y p_{n-2} son casi iguales.

17. Se necesitan siete iteraciones.

19. Para $f(x) = \ln(x^2 + 1) - e^{0.4x} \cos \pi x$, tenemos las siguientes raíces.

a. Para $p_0 = -0.5$, tenemos $p_3 = 0.4341431$.

b. Para $p_0 = 0.5$, tenemos $p_3 = 0.4506567$.

Para $p_0 = 1.5$, tenemos $p_3 = 1.7447381$.

Para $p_0 = 2.5$, tenemos $p_5 = 2.2383198$.

Para $p_0 = 3.5$, tenemos $p_4 = 3.7090412$.

c. La aproximación inicial $n - 0.5$ es muy razonable.

d. Para $p_0 = 24.5$, tenemos $p_2 = 24.4998870$.

21. Los dos números son aproximadamente 6.512849 y 13.487151.

23. El solicitante de crédito está en condiciones de pagar máximo un 8.10 %.

25. a. `solve(3^(3*x+1)-7*5^(2*x),x)` y `fsolve(3^(3*x+1)-7*5^(2*x),x)` fallan.

b. `plot(3^(3*x+1)-7*5^(2*x),x=a..b)` generalmente no produce información útil. Sin embargo, $a = 10.5$ y $b = 11.5$ en el comando de graficación muestran que $f(x)$ tiene una raíz cercana a $x = 11$.

c. Con $p_0 = 11$, $p_5 = 11.0094386442681716$ tiene una exactitud de 10^{-16}.

d. $p = \dfrac{\ln(3/7)}{\ln(25/27)}$

27. Tenemos $P_L = 265816$, $c = -0.75658125$ y $k = 0.045017502$. La población de 1980 es $P(30) = 222\,248\,320$, y la del año 2010 es $P(60) = 252\,967\,030$.

29. Al usar $p_0 = 0.5$ y $p_1 = 0.9$, el método de la secante da $p_5 = 0.842$.

Conjunto de ejercicios 2.4

1. a. Para $p_0 = 0.5$, tenemos $p_{13} = 0.567135$.

b. Para $p_0 = -1.5$, tenemos $p_{23} = -1.414325$.

c. Para $p_0 = 0.5$, tenemos $p_{22} = 0.641166$.

d. Para $p_0 = -0.5$, tenemos $p_{23} = -0.183274$.

3. El método de Newton con $p_0 = -0.5$ produce $p_{13} = -0.169607$. El método modificado de Newton en la ecuación (2.11) con $p_0 = -0.5$ produce $p_{11} = -0.169607$.

5. a. Con $k > 0$,

$$\lim_{n\to\infty} \frac{|p_{n+1} - 0|}{|p_n - 0|} = \lim_{n\to\infty} \frac{\frac{1}{(n+1)^k}}{\frac{1}{n^k}} = \lim_{n\to\infty}\left(\frac{n}{n+1}\right)^k = 1,$$

y, por tanto, la convergencia es lineal.

b. Necesitamos tener $N > 10^{m/k}$.

7. Ejemplos típicos son **a.** $p_n = 10^{-3^n}$ y **b.** $p_n = 10^{-\alpha^n}$.

9. Esto se deduce del hecho de que

$$\lim_{n\to\infty} \frac{\left|\frac{b-a}{2^{n+1}}\right|}{\left|\frac{b-a}{2^n}\right|} = \frac{1}{2}.$$

11. Si $\frac{|p_{n+1}-p|}{|p_n-p|^2} = 0.75$ y $|p_0 - p| = 0.5$, entonces

$$|p_n - p| = (0.75)^{(2^n-1)}|p_0 - p|^{2^n}.$$

Para tener $|p_n - p| \le 10^{-8}$, se requiere que $n \ge 3$.

Conjunto de ejercicios 2.5

1. Los resultados se incluyen en la tabla anexa.

	a	b	c	d
$\hat{p}_0$	0.258684	0.907859	0.548101	0.731385
$\hat{p}_1$	0.257613	0.909568	0.547915	0.736087
$\hat{p}_2$	0.257536	0.909917	0.547847	0.737653
$\hat{p}_3$	0.257531	0.909989	0.547823	0.738469
$\hat{p}_4$	0.257530	0.910004	0.547814	0.738798
$\hat{p}_5$	0.257530	0.910007	0.547810	0.738958

3. $p_0^{(1)} = 0.826427$

5. $p_1^{(0)} = 1.5$

7. Para $g(x) = \sqrt{1 + \frac{1}{x}}$ y $p_0 = 1$, tenemos $p_3 = 1.32472$.

9. Para $g(x) = 0.5(x + \frac{3}{x})$ y $p_0 = 0.5$, tenemos $p_4 = 1.73205$.

11. El método Δ^2 de Aitken produce: **a.** $\hat{p}_{10} = 0.0\overline{45}$ **b.** $\hat{p}_2 = 0.0363$

13. Tenemos

$$\frac{|p_{n+1} - p_n|}{|p_n - p|} = \frac{|p_{n+1} - p + p - p_n|}{|p_n - p|} = \left|\frac{p_{n+1} - p}{p_n - p} - 1\right|,$$

por tanto,

$$\lim_{n\to\infty} \frac{|p_{n+1} - p_n|}{|p_n - p|} = \lim_{n\to\infty}\left|\frac{p_{n+1} - p}{p_n - p} - 1\right| = 1.$$

15. a. *Sugerencia:* Demuestre primero que $p_n - p = -\frac{1}{(n+1)!}e^{\xi}x^{n+1}$, donde ξ está entre 0 y 1.

b.

n	p_n	$\hat{p}_n$
0	1	3
1	2	2.75
2	2.5	$2.7\overline{2}$
3	$2.\overline{6}$	2.71875
4	$2.708\overline{3}$	$2.718\overline{3}$
5	$2.71\overline{6}$	2.7182870
6	$2.7180\overline{5}$	2.7182823
7	2.7182539	2.7182818
8	2.7182787	2.7182818
9	2.7182815	
10	2.7182818	

Conjunto de ejercicios 2.6

1. a. Para $p_0 = 1$, tenemos $p_{22} = 2.69065$.

b. Para $p_0 = 1$, tenemos $p_5 = 0.53209$; para $p_0 = 1$, tenemos $p_3 = -0.65270$; y para $p_0 = -3$, tenemos $p_3 = -2.87939$.

c. Para $p_0 = 1$, tenemos $p_5 = 1.32472$.

d. Para $p_0 = 1$, tenemos $p_4 = 1.12412$; y para $p_0 = 0$, tenemos $p_8 = -0.87605$.

e. Para $p_0 = 0$, tenemos $p_6 = -0.47006$; para $p_0 = -1$, tenemos $p_4 = -0.88533$, y para $p_0 = -3$, tenemos $p_4 = -2.64561$.

f. Para $p_0 = 0$, tenemos $p_{10} = 1.49819$.

3. La tabla adjunta contiene la aproximación inicial y las raíces.

	p_0	p_1	p_2	Raíces aproximadas	Raíces combinadas complejas
a.	-1	0	1	$p_7 = -0.34532 - 1.31873i$	$-0.34532 + 1.31873i$
	0	1	2	$p_6 = 2.69065$	
b.	0	1	2	$p_6 = 0.53209$	
	1	2	3	$p_9 = -0.65270$	
	-2	-3	-2.5	$p_4 = -2.87939$	
c.	0	1	2	$p_5 = 1.32472$	
	-2	-1	0	$p_7 = -0.66236 - 0.56228i$	$-0.66236 + 0.56228i$
d.	0	1	2	$p_5 = 1.12412$	
	2	3	4	$p_{12} = -0.12403 + 1.74096i$	$-0.12403 - 1.74096i$
	-2	0	-1	$p_5 = -0.87605$	
e.	0	1	2	$p_{10} = -0.88533$	
	1	0	-0.5	$p_5 = -0.47006$	
	-1	-2	-3	$p_5 = -2.64561$	
f.	0	1	2	$p_6 = 1.49819$	
	-1	-2	-3	$p_{10} = -0.51363 - 1.09156i$	$-0.51363 + 1.09156i$
	1	0	-1	$p_8 = 0.26454 - 1.32837i$	$0.26454 + 1.32837i$

5. a. Las raíces son 1.244, 8.847 y -1.091, y los puntos críticos son 0 y 6.

b. Las raíces son 0.5798, 1.521, 2.332 y -2.432, y los puntos críticos son 1, 2.001 y -1.5.

7. Sean $c_1 = (2 + \frac{2}{9}\sqrt{129})^{-1/3}$ y $c_2 = (2 + \frac{2}{9}\sqrt{129})^{1/3}$. Las raíces son $c_2 - \frac{4}{3}c_1$, $-\frac{1}{2}c_2 + \frac{2}{3}c_1 + \frac{1}{2}\sqrt{3}\,(c_2 + \frac{4}{3}c_1)i$, y $-\frac{1}{2}c_2 + \frac{2}{3}c_1 - \frac{1}{2}\sqrt{3}\,(c_2 + \frac{4}{3}c_1)i$.

9. Todos los métodos encuentran la solución 0.23235.

11. El material mínimo aproximadamente es 573.64895 cm^2.

Conjunto de ejercicios 3.1

1. a. $P_1(x) = -0.148878x + 1$; $P_2(x) = -0.452592x^2 - 0.0131009x + 1$; $P_1(0.45) = 0.933005$; $|f(0.45) - P_1(0.45)| = 0.032558$; $P_2(0.45) = 0.902455$; $|f(0.45) - P_2(0.45)| = 0.002008$

b. $P_1(x) = 0.467251x + 1$; $P_2(x) = -0.0780026x^2 + 0.490652x + 1$; $P_1(0.45) = 1.210263$; $|f(0.45) - P_1(0.45)| = 0.006104$; $P_2(0.45) = 1.204998$; $|f(0.45) - P_2(0.45)| = 0.000839$

c. $P_1(x) = 0.874548x$; $P_2(x) = -0.268961x^2 + 0.955236x$; $P_1(0.45) = 0.393546$; $|f(0.45) - P_1(0.45)| = 0.0212983$; $P_2(0.45) = 0.375392$; $|f(0.45) - P_2(0.45)| = 0.003828$

d. $P_1(x) = 1.031121x$; $P_2(x) = 0.615092x^2 + 0.846593x$; $P_1(0.45) = 0.464004$; $|f(0.45) - P_1(0.45)| = 0.019051$; $P_2(0.45) = 0.505523$; $|f(0.45) - P_2(0.45)| = 0.022468$

3. a.

n	$x_0, x_1, \ldots, x_n$	$P_n(8.4)$
1	8.3, 8.6	17.87833
2	8.3, 8.6, 8.7	17.87716
3	8.3, 8.6, 8.7, 8.1	17.87714

b.

n	$x_0, x_1, \ldots, x_n$	$P_n(-1/3)$
1	−0.5, −0.25	0.21504167
2	−0.5, −0.25, 0.0	0.16988889
3	−0.5, −0.25, 0.0, −0.75	0.17451852

c.

n	$x_0, x_1, \ldots, x_n$	$P_n(0.25)$
1	0.2, 0.3	−0.13869287
2	0.2, 0.3, 0.4	−0.13259734
3	0.2, 0.3, 0.4, 0.1	−0.13277477

d.

n	$x_0, x_1, \ldots, x_n$	$P_n(0.9)$
1	0.8, 1.0	0.44086280
2	0.8, 1.0, 0.7	0.43841352
3	0.8, 1.0, 0.7, 0.6	0.44198500

5. $\sqrt{3} \approx P_4(\frac{1}{2}) = 1.708\overline{3}$

7. a.

n	Error real	Cota de error
1	0.00118	0.00120
2	1.367×10^{-5}	1.452×10^{-5}

b.

n	Error real	Cota de error
1	4.0523×10^{-2}	4.5153×10^{-2}
2	4.6296×10^{-3}	4.6296×10^{-3}

c.

n	Error real	Cota de error
1	5.9210×10^{-3}	6.0971×10^{-3}
2	1.7455×10^{-4}	1.8128×10^{-4}

d.

n	Error real	Cota de error
1	2.7296×10^{-3}	1.4080×10^{-2}
2	5.1789×10^{-3}	9.2215×10^{-3}

9. $y = 4.25$.

11. $f(1.09) \approx 0.2826$. El error real es 4.3×10^{-5}, y una cota de error es 7.4×10^{-6}. La discrepancia se debe al hecho de que los datos se dan sólo en cuatro cifras decimales y se emplea exclusivamente una aritmética de cuatro dígitos.

13. $P_2 = f(0.7) = 6.4$.

15. a. $P_2(x) = -11.22388889x^2 + 3.810500000x + 1$, y una cota de error es 0.11371294.

b. $P_2(x) = -0.1306344167x^2 + 0.8969979335x - 0.63249693$, y una cota de error es 9.45762×10^{-4}.

c. $P_3(x) = 0.1970056667x^3 - 1.06259055x^2 + 2.532453189x - 1.666868305$, y una cota de error es 10^{-4}.

d. $P_3(x) = -0.07932x^3 - 0.545506x^2 + 1.0065992x + 1$, y una cota de error es 1.591376×10^{-3}.

17. El mayor tamaño posible de paso es 0.004291932, así que 0.04 sería una elección razonable.

19. $P_{0,1,2,3}(2.5) = 2.875$.

21. Los primeros diez términos de la sucesión son 0.038462, 0.0333671, 0.116605, −0.371760, −0.0548919, 0.605935, 0.190249, −0.513353, −0.0668173 y 0.448335. Puesto que $f(1+\sqrt{10}) = 0.0545716$, la sucesión no parece converger.

25. a. Muestra 1:

$P_6(x) = 6.67 - 42.6434x + 16.1427x^2 - 2.09464x^3 + 0.126902x^4 - 0.00367168x^5 + 0.0000409458x^6$; Muestra 2:

$P_6(x) = 6.67 - 5.67821x + 2.91281x^2 - 0.413799x^3 + 0.0258413x^4 - 0.000752546x^5 + 0.00000836160x^6$

b. Muestra 1: 42.71 mg; Muestra 2: 19.42 mg

27. Puesto que $g(x) = g(x_0) = 0$, existe un número ξ_1 entre x y x_0 para el cual $g'(\xi_1) = 0$. Por lo demás, $g'(x_0) = 0$, de modo que existe un número ξ_2 entre x_0 y ξ_1 para el cual $g''(\xi_2) = 0$. Se sigue realizando el proceso por inducción para demostrar que existe un número ξ_{n+1} entre x_0 y ξ_n con $g^{(n+1)}(\xi_{n+1}) = 0$. A continuación se transcribe la fórmula de error para los polinomios de Taylor.

29. a. (i) $B_3(x) = x$ (ii) $B_3(x) = 1$

Conjunto de ejercicios 3.2

1. a. $P_1(x) = 16.9441 + 3.1041(x - 8.1)$; $P_1(8.4) = 17.87533$; $P_2(x) = P_1(x) + 0.06(x - 8.1)(x - 8.3)$; $P_2(8.4) = 17.87713$; $P_3(x) = P_2(x) + -0.00208333(x - 8.1)(x - 8.3)(x - 8.6)$; $P_3(8.4) = 17.87714$

b. $P_1(x) = -0.1769446 + 1.9069687(x - 0.6)$; $P_1(0.9) = 0.395146$; $P_2(x) = P_1(x) + 0.959224(x - 0.6)(x - 0.7)$; $P_2(0.9) = 0.4526995$; $P_3(x) = P_2(x) - 1.785741(x - 0.6)(x - 0.7)(x - 0.8)$; $P_3(0.9) = 0.4419850$

3. En las siguientes ecuaciones tenemos $s = \frac{1}{h}(x - x_n)$.

a. $P_1(s) = 1.101 + 0.7660625s$; $f(-\frac{1}{3}) \approx P_1(-\frac{4}{3}) = 0.07958333$

$P_2(s) = P_1(s) + 0.406375s(s+1)/2$; $f(-\frac{1}{3}) \approx P_2(-\frac{4}{3}) = 0.1698889$

$P_3(s) = P_2(s) + 0.09375s(s+1)(s+2)/6$; $f(-\frac{1}{3}) \approx P_3(-\frac{4}{3}) = 0.1745185$

b. $P_1(s) = 0.2484244 + 0.2418235s$; $f(0.25) \approx P_1(-1.5) = -0.1143108$; $P_2(s) = P_1(s) - 0.04876419s(s+1)/2$; $f(0.25) \approx P_2(-1.5) = -0.1325973$; $P_3(s) = P_2(s) - 0.00283891s(s+1)(s+2)/6$; $f(0.25) \approx P_3(-1.5) = -0.1327748$

5. a. $f(0.05) \approx 1.05126$ **b.** $f(0.65) \approx 1.91555$ **c.** $f(0.43) \approx 1.53725$

7. a. $P(-2) = Q(-2) = -1$, $P(-1) = Q(-1) = 3$, $P(0) = Q(0) = 1$, $P(1) = Q(1) = -1$, $P(2) = Q(2) = 3$

b. El formato del polinomio no es único. Si ampliamos $P(x)$ y $Q(x)$, son idénticos. Sólo hay un polinomio interpolante de grado menor o igual que los cuatro para los datos. Pero puede expresarse de varias maneras, dependiendo de la aplicación.

9. El coeficiente de x^2 es 3.5.

11. La aproximación a $f(0.3)$ debería aumentarse en 5.9375.

13. $f[x_0] = f[x_0] = 1$, $f[x_1] = f(x_1) = 3$, $f[x_0, x_1] = 5$

15. Puesto que $f[x_2] = f[x_0] + f[x_0, x_1](x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$,

$$a_2 = \frac{f[x_2] - f[x_0]}{(x_2 - x_0)(x_2 - x_1)} - \frac{f[x_0, x_1]}{(x_2 - x_1)}.$$

Esto se simplifica y se convierte en $f[x_0, x_1, x_2]$.

17. Sea $\tilde{P}(x) = f[x_{i_0}] + \sum_{k=1}^{n} f[x_{i_0}, \ldots, x_{i_k}](x - x_{i_0}) \cdots (x - x_{i_{k-1}})$ y $\hat{P}(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, \ldots, x_k](x - x_0) \cdots (x - x_{k-1})$. El polinomio $\tilde{P}(x)$ interpola $f(x)$ en los nodos $x_{i_0}, \ldots, x_{i_n}$ y el polinomio $\hat{P}(x)$ interpola $f(x)$ en los nodos $x_0, \ldots, x_n$. Como ambos conjuntos de nodos son los mismos y el polinomio interpolante es único, tenemos $\tilde{P}(x) = \hat{P}(x)$. El coeficiente de x^n en $\tilde{P}(x)$ es $f[x_{i_0}, \ldots, x_{i_n}]$ y el coeficiente de x^n en $\hat{P}(x)$ es $f[x_0, \ldots, x_n]$. Por tanto, $f[x_{i_0}, \ldots, x_{i_n}] = f[x_0, \ldots, x_n]$.

Conjunto de ejercicios 3.3

1. En las tablas anexas se incluyen los coeficientes de los polinomios en forma de diferencias divididas. Por ejemplo, el polinomio del inciso (a) es

$$H_3(x) = 17.56492 + 3.116256(x - 8.3) + 0.05948(x - 8.3)^2 - 0.00202222(x - 8.3)^2(x - 8.6).$$

a	b	c	d
17.56492	0.022363362	−0.02475	−0.62049958
3.116256	2.1691753	0.751	3.5850208
0.05948	0.01558225	2.751	−2.1989182
−0.00202222	−3.2177925	1	−0.490447
		0	0.037205
		0	0.040475
			−0.0025277777
			0.0029629628

3. **a.** Tenemos sen $0.34 \approx H_5(0.34) = 0.33349$.

b. La fórmula da una cota de error de 3.05×10^{-14}, pero el error real es 2.91×10^{-6}. La discrepancia se debe al hecho de que los datos contienen sólo cinco cifras decimales.

c. Tenemos sen $0.34 \approx H_7(0.34) = 0.33350$. Aunque ahora la cota de error es 5.4×10^{-20}, la exactitud de los datos domina los cálculos. En realidad el resultado es menos preciso que la aproximación del inciso (b) porque sen $0.34 = 0.333487$.

5. En 2(a) tenemos una cota de error de 5.9×10^{-8}. La cota de error en 2 (c) es 0 porque $f^{(n)}(x) \equiv 0$ para $n > 3$.

7. El polinomio de Hermite generado de estos datos es

$$\begin{aligned} H_9(x) = {} & 75x + 0.222222x^2(x - 3) - 0.0311111x^2(x - 3)^2 - 0.00644444x^2(x - 3)^2(x - 5) \\ & + 0.00226389x^2(x - 3)^2(x - 5)^2 - 0.000913194x^2(x - 3)^2(x - 5)^2(x - 8) \\ & + 0.000130527x^2(x - 3)^2(x - 5)^2(x - 8)^2 - 0.0000202236x^2(x - 3)^2(x - 5)^2(x - 8)^2(x - 13). \end{aligned}$$

a. El polinomio de Hermite predice una posición de $H_9(10) = 743$ pies y una rapidez de $H_9'(10) = 48$ pies/s. Aunque la aproximación de la posición es razonable, la predicción de baja rapidez provoca sospechas.

b. Para calcular la primera vez que la rapidez rebasa los 55 mi/h $= 80.\overline{6}$ pie/s, resolvemos el valor más pequeño de t en la ecuación $80.\overline{6} = H_9'(x)$. Esto nos da $x \approx 5.6488092$.

c. La rapidez máxima estimada es de $H_9'(x)(12.37187) = 119.423$ pie/s ≈ 81.425 mi/h.

9. Sea

$$H(x) = f[z_0] + f[z_0, z_1](x - x_0) + f[z_0, z_1, z_2](x - x_0)^2 + f[z_0, z_1, z_2, z_3](x - x_0)^2(x - x_1).$$

Al sustituir $f[z_0] = f(x_0)$, $f[z_0, z_1] = f'(x_0)$,

$$f[z_0, z_1, z_2] = \frac{f(x_1) - f(x_0) - f'(x_0)(x_1 - x_0)}{x_1 - x_0}$$

$$f[z_0, z_1, z_2, z_3] = \frac{f'(x_1)(x_1 - x_0) - 2f(x_1) + 2f(x_0) + f'(x_0)(x_1 - x_0)}{(x_1 - x_0)^3}$$

en $H(x)$ y, al simplificar, obtenemos

$$H(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f(x_1) - f(x_0) - f'(x_0)(x_1 - x_0)}{(x_1 - x_0)^2}(x - x_0)^2$$
$$+ \frac{f'(x_1)(x_1 - x_0) - 2f(x_1) + 2f(x_0) + f'(x_0)(x_1 - x_0)}{(x_1 - x_0)^3}(x - x_0)^2(x - x_1).$$

Por tanto, $H(x_0) = f(x_0)$ y

$$H(x_1) = f(x_0) + f'(x_0)(x_1 - x_0) + [f(x_1) - f(x_0) - f'(x_0)(x_1 - x_0)]$$
$$= f(x_1).$$

Más aún,

$$H'(x) = f'(x_0) + 2\,\frac{f(x_1) - f(x_0) - f'(x_0)(x_1 - x_0)}{(x_1 - x_0)^2}(x - x_0)$$
$$+ \frac{f'(x_1)(x_1 - x_0) - 2f(x_1) + 2f(x_0) + f'(x_0)(x_1 - x_0)}{(x_1 - x_0)^3}[2(x - x_0)(x - x_1) + (x - x_0)^2].$$

así que

$$H'(x_0) = f'(x_0) \quad \text{y}$$
$$H'(x_1) = f'(x_0) + \frac{2f(x_1)}{x_1 - x_0} - \frac{2f(x_0)}{x_1 - x_0} - 2f'(x_0) + f'(x_1) - \frac{2f(x_1)}{x_1 - x_0} + \frac{2f(x_0)}{x_1 - x_0} + f'(x_0)$$
$$= f'(x_1).$$

Por tanto H, cumple los requisitos del polinomio cúbico de Hermite H_3, y la singularidad de H_3 implica que $H_3 = H$.

Conjunto de ejercicios 3.4

1. $S(x) = x$ en $[0, 2]$
3. Las ecuaciones de los respectivos trazadores cúbicos libres están dadas por

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

para x en $[x_i, x_{i+1}]$: los coeficientes se proporcionan en las tablas siguientes.

a.

i	a_i	b_i	c_i	d_i
0	17.564920	3.13410000	0.00000000	0.00000000

b.

i	a_i	b_i	c_i	d_i
0	0.22363362	2.17229175	0.00000000	0.00000000

c.

i	a_i	b_i	c_i	d_i
0	−0.02475000	1.03237500	0.00000000	6.50200000
1	0.33493750	2.25150000	4.87650000	−6.50200000

d.

i	a_i	b_i	c_i	d_i
0	−0.62049958	3.45508693	0.00000000	−8.9957933
1	−0.28398668	3.18521313	−2.69873800	−0.94630333
2	0.00660095	2.61707643	−2.98262900	9.9420966

5. Las ecuaciones de los respectivos trazadores cúbicos sujetos están dadas por

$$s(x) = s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

para x en $[x_i, x_{i+1}]$; los coeficientes se incluyen en las tablas anexas.

a.

i	a_i	b_i	c_i	d_i
0	17.564920	3.1162560	0.0600867	−0.00202222

b.

i	a_i	b_i	c_i	d_i
0	0.22363362	2.1691753	0.65914075	−3.2177925

c.

i	a_i	b_i	c_i	d_i
0	−0.02475000	0.75100000	2.5010000	1.000000
1	0.33493750	2.18900000	3.2510000	1.000000

d.

i	a_i	b_i	c_i	d_i
0	−0.62049958	3.5850208	−2.1498407	−0.49077413
1	−0.28398668	3.1403294	−2.2970730	−0.47458360
2	0.006600950	2.6666773	−2.4394481	−0.44980146

7. $b = -1$, $c = -3$, $d = 1$

9. $B = \frac{1}{4}$, $D = \frac{1}{4}$, $b = -\frac{1}{2}$, $d = \frac{1}{4}$

11. La ecuación del trazador es

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

en el intervalo $[x_i, x_{i+1}]$; los coeficientes se incluyen en la tabla anexa.

x_i	a_i	b_i	c_i	d_i
0	1.0	−0.7573593	0.0	−6.627417
0.25	0.7071068	−2.0	−4.970563	6.627417
0.5	0.0	−3.242641	0.0	6.627417
0.75	−0.7071068	−2.0	4.970563	−6.627417

$\int_0^1 S(x)dx = 0.000000$, $S'(0.5) = -3.24264$ y $S''(0.5) = 0.0$

13. La ecuación del trazador es

$$s(x) = s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

en el intervalo $[x_i, x_{i+1}]$; los coeficientes se incluyen en la tabla anexa.

x_i	a_i	b_i	c_i	d_i
0	1.0	0.0	−5.193321	2.028118
0.25	0.7071068	−2.216388	−3.672233	4.896310
0.5	0.0	−3.134447	0.0	4.896310
0.75	−0.7071068	−2.216388	3.672233	2.028118

$\int_0^1 s(x)\,dx = 0.000000$, $s'(0.5) = -3.13445$ y $s''(0.5) = 0.0$

15. Sea $f(x) = a + bx + cx^2 + dx^3$. Es claro que f satisface las propiedades (a), (c), (d) y (e) de la definición 3.10 y que f se interpola para cualquier elección de $x_0, \ldots, x_n$. Como se cumple (ii) de (f) en la definición 3.10, f debe ser su propio trazador cúbico sujeto. Sin embargo, $f''(x) = 2c + 6dx$ puede ser cero sólo en $x = -c/3d$. En consecuencia, el inciso (i) de (f) en la definición 3.10 no puede sostenerse en dos valores x_0 y x_n. Por consiguiente, f tampoco puede ser un trazador cúbico natural.

17. La aproximación lineal segmentaria de f está dada por

$$F(x) = \begin{cases} 20(e^{0.1} - 1)x + 1 & \text{para } x \text{ en } [0, 0.05] \\ 20(e^{0.2} - e^{0.1})x + 2e^{0.1} - e^{0.2}, & \text{para } x \text{ en } (0.05, 1]. \end{cases}$$

Tenemos

$$\int_0^{0.1} F(x)\,dx = 0.1107936 \quad \text{y} \quad \int_0^{0.1} f(x)\,dx = 0.1107014$$

21. a. En $[0, 0.05]$ tenemos $s(x) = 1.000000 + 1.999999x + 1.998302x^2 + 1.401310x^3$ y en $(0.05, 0.1]$ tenemos $s(x) = 1.105170 + 2.210340(x - 0.05) + 2.208498(x - 0.05)^2 + 1.548758(x - 0.05)^3$.

b. $\int_0^{0.1} s(x)\,dx = 0.110701$

c. 1.6×10^{-7}

d. En $[0, 0.05]$ tenemos $S(x) = 1 + 2.04811x + 22.12184x^3$ y en $(0.05, 0.1]$ tenemos $S(x) = 1.105171 + 2.214028(x - 0.05) + 3.318277(x - 0.05)^2 - 22.12184(x - 0.05)^3$. $S(0.02) = 1.041139$ y $S(0.02) = 1.040811$.

23.

$$S(x) = \begin{cases} 2x - x^2, & 0 \le x \le 1 \\ 1 + (x - 1)^2, & 1 \le x \le 2 \end{cases}$$

25. El trazador tiene la ecuación

$$s(x) = s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

para x en $[x_i, x_{i+1}]$; los coeficientes se incluyen en la tabla anexa.

x_i	a_i	b_i	c_i	d_i
0	0	75	−0.659292	0.219764
3	225	76.9779	1.31858	−0.153761
5	383	80.4071	0.396018	−0.177237
8	623	77.9978	−1.19912	0.0799115

El trazador predice una posición de $s(10) = 774.84$ pies y una rapidez de $s'(10) = 74.16$ pies/s. Para aumentar al máximo la rapidez, calculamos los puntos críticos de $s'(x)$ y después comparamos los valores de $s(x)$ en estos puntos y en los extremos. Encontramos que máx $s'(x) = s'(5.7448) = 80.7$ pies/s $= 55.02$ mi/h. La rapidez de 55 mi/h fue rebasada primero en aproximadamente 5.5 s.

27. La ecuación del trazador es

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

para x en el intervalo $[x_i, x_{i+1}]$; los coeficientes se incluyen en la tabla anexa.

	Ejemplo 1				Ejemplo 2			
x_i	a_i	b_i	c_i	d_i	a_i	b_i	c_i	d_i
0	6.67	−0.44687	0	0.06176	6.67	1.6629	0	−0.00249
6	17.33	6.2237	1.1118	−0.27099	16.11	1.3943	−0.04477	−0.03251
10	42.67	2.1104	−2.1401	0.28109	18.89	−0.52442	−0.43490	0.05916
13	37.33	−3.1406	0.38974	−0.01411	15.00	−1.5365	0.09756	0.00226
17	30.10	−0.70021	0.22036	−0.02491	10.56	−0.64732	0.12473	−0.01113
20	29.31	−0.05069	−0.00386	0.00016	9.44	−0.19955	0.02453	−0.00102

29. Los tres trazadores naturales tienen ecuaciones de la forma

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

para x en $[x_i, x_{i+1}]$; los valores de los coeficientes se incluyen en las tablas adjuntas.

Trazador 1

i	x_i	$a_i = f(x_i)$	b_i	c_i	d_i
0	1	3.0	0.786	0.0	−0.086
1	2	3.7	0.529	−0.257	0.034
2	5	3.9	−0.086	0.052	0.334
3	6	4.2	1.019	1.053	−0.572
4	7	5.7	1.408	−0.664	0.156
5	8	6.6	0.547	−0.197	0.024
6	10	7.1	0.049	−0.052	−0.003
7	13	6.7	−0.342	−0.078	0.007
8	17	4.5			

Trazador 2

i	x_i	$a_i = f(x_i)$	b_i	c_i	d_i
0	17	4.5	1.106	0.0	−0.030
1	20	7.0	0.289	−0.272	0.025
2	23	6.1	−0.660	−0.044	0.204
3	24	5.6	−0.137	0.567	−0.230
4	25	5.8	0.306	−0.124	−0.089
5	27	5.2	−1.263	−0.660	0.314
6	27.7	4.1			

Trazador 3

i	x_i	$a_i = f(x_i)$	b_i	c_i	d_i
0	27.7	4.1	0.749	0.0	−0.910
1	28	4.3	0.503	−0.819	0.116
2	29	4.1	−0.787	−0.470	0.157
3	30	3.0			

Conjunto de ejercicios 3.5

1. **a.** $x(t) = -10t^3 + 14t^2 + t, \quad y(t) = -2t^3 + 3t^2 + t$

b. $x(t) = -10t^3 + 14.5t^2 + 0.5t, \quad y(t) = -3t^3 + 4.5t^2 + 0.5t$

c. $x(t) = -10t^3 + 14t^2 + t, \quad y(t) = -4t^3 + 5t^2 + t$

d. $x(t) = -10t^3 + 13t^2 + 2t, \quad y(t) = 2t$

3. **a.** $x(t) = -11.5t^3 + 15t^2 + 1.5t, +1, \quad y(t) = -4.25t^3 + 4.5t^2 + 0.75t + 1$

b. $x(t) = -6.25t^3 + 10.5t^2 + 0.75t + 1, \quad y(t) = -3.5t^3 + 3t^2 + 1.5t + 1$

c. Para t entre (0, 0) y (4, 6) tenemos

$$x(t) = -5t^3 + 7.5t^2 + 1.5t, \quad y(t) = -13.5t^3 + 18t^2 + 1.5t,$$

y para t entre (4, 6) y (6, 1) tenemos

$$x(t) = -5.5t^3 + 6t^2 + 1.5t + 4, \quad y(t) = 4t^3 - 6t^2 - 3t + 6.$$

d. Para t entre (0, 0) y (2, 1) tenemos

$$x(t) = -5.5t^3 + 6t^2 + 1.5t, \quad y(t) = -0.5t^3 + 1.5t,$$

para t entre (2, 1) y (4, 0) tenemos

$$x(t) = -4t^3 + 3t^2 + 3t + 2, \quad y(t) = -t^3 + 1,$$

y para t entre (4, 0) y (6, −1) tenemos

$$x(t) = -8.5t^3 + 13.5t^2 - 3t + 4, \quad y(t) = -3.25t^3 + 5.25t^2 - 3t.$$

Conjunto de ejercicios 4.1

1. Conforme a la fórmula de diferencias progresivas-regresivas (4.1), tenemos las siguientes aproximaciones:

a. $f'(0.5) \approx 0.8520, f'(0.6) \approx 0.8520, f'(0.7) \approx 0.7960$

b. $f'(0.0) \approx 3.7070, f'(0.2) \approx 3.1520, f'(0.4) \approx 3.1520$

3. En el caso de los extremos de las tablas utilizamos la fórmula (4.4). Las otras aproximaciones provienen de la fórmula (4.5).

a. $f'(1.1) \approx 17.769705, f'(1.2) \approx 22.193635, f'(1.3) \approx 27.107350, \; f'(1.4) \approx 32.150850$

b. $f'(8.1) \approx 3.092050, f'(8.3) \approx 3.116150, f'(8.5) \approx 3.139975, \; f'(8.7) \approx 3.163525$

c. $f'(2.9) \approx 5.101375, f'(3.0) \approx 6.654785, f'(3.1) \approx 8.216330, \; f'(3.2) \approx 9.786010$

d. $f'(2.0) \approx 0.13533150, f'(2.1) \approx -0.09989550, f'(2.2) \approx -0.3298960, \; f'(2.3) \approx -0.5546700$

5. Las aproximaciones y las fórmulas usadas son:

a. $f'(2.1) \approx 3.899344$ de (4.7) $\quad f'(2.2) \approx 2.876876$ de (4.7) $\quad f'(2.3) \approx 2.249704$ de (4.6)

$f'(2.4) \approx 1.837756$ de (4.6) $\quad f'(2.5) \approx 1.544210$ de (4.7) $\quad f'(2.6) \approx 1.355496$ de (4.7)

b. $f'(-3.0) \approx -5.877358$ de (4.7) $\quad f'(-2.8) \approx -5.468933$ de (4.7) $\quad f'(-2.6) \approx -5.059884$ de (4.6)

$f'(-2.4) \approx -4.650223$ de (4.6) $\quad f'(-2.2) \approx -4.239911$ de (4.7) $\quad f'(-2.0) \approx -3.828853$ de (4.7)

7. $f'(3) \approx \frac{1}{12}[f(1) - 8f(2) + 8f(4) - f(5)] = 0.21062$, con una cota de error dada por

$$\max_{1\le x\le 5} \frac{|f^{(5)}(x)|\,h^4}{30} \le \frac{23}{30} = 0.7\bar{6}.$$

9. Conforme a la fórmula progresiva-regresiva (4.1), tenemos las siguientes aproximaciones:

a. $f'(0.5) \approx 0.852, f'(0.6) \approx 0.852, f'(0.7) \approx 0.7960$

b. $f'(0.0) \approx 3.707, f'(0.2) \approx 3.153, f'(0.4) \approx 3.153$

11. En el caso de los extremos de las tablas utilizamos la fórmula (4.7). Las otras aproximaciones provienen de la fórmula (4.6).

a. $f'(2.1) \approx 3.884$ $f'(2.2) \approx 2.896$ $f'(2.3) \approx 2.249$ $f'(2.4) \approx 1.836$ $f'(2.5) \approx 1.550$ $f'(2.6) \approx 1.348$

b. $f'(-3.0) \approx -5.883$ $f'(-2.8) \approx -5.467$ $f'(-2.6) \approx -5.059$ $f'(-2.4) \approx -4.650$ $f'(-2.2) \approx -4.208$ $f'(-2.0) \approx -3.875$

13. La aproximación es -4.8×10^{-9}. $f''(0.5) = 0$. La cota de error es 0.35874. Los métodos son muy exactos porque la función es simétrica alrededor de $x = 0.5$.

15. **a.** $f'(0.2) \approx -0.1951027$ **b.** $f'(1.0) \approx -1.541415$ **c.** $f'(0.6) \approx -0.6824175$

17. $f'(0.4) \approx -0.4249840$ y $f'(0.8) \approx -1.032772$.

19. Al utilizar la fórmula de tres puntos, obtenemos la siguiente tabla:

Tiempo	0	3	5	8	10	13
Rapidez	79	82.4	74.2	76.8	69.4	71.2

21. Las aproximaciones finalmente se convierten en cero, porque el numerador se transforma en cero.

23. Dado que $e'(h) = -\varepsilon/h^2 + hM/3$, tenemos $e'(h) = 0$ si y sólo si $h = \sqrt[3]{3\varepsilon/M}$. Por otra parte, $e'(h) < 0$ si $h < \sqrt[3]{3\varepsilon/M}$ y $e'(h) > 0$ si $h > \sqrt[3]{3\varepsilon/M}$, así que un mínimo absoluto de $e(h)$ ocurre en $h = \sqrt[3]{3\varepsilon/M}$.

Conjunto de ejercicios 4.2

1. **a.** $f'(1) \approx 1.0000109$ **b.** $f'(0) \approx 2.0000000$ **c.** $f'(1.05) \approx 2.2751459$ **d.** $f'(2.3) \approx -19.646799$

3. **a.** $f'(1) \approx 1.001$ **b.** $f'(0) \approx 1.999$ **c.** $f'(1.05) \approx 2.283$ **d.** $f'(2.3) \approx -19.61$

5. $\int_0^{\pi} \operatorname{sen} x \, dx \approx 1.999999$

9. Sea $N_2(h) = N\left(\frac{h}{3}\right) + \left(\frac{N\left(\frac{h}{3}\right) - N(h)}{2}\right)$ y $N_3(h) = N_2\left(\frac{h}{3}\right) + \left(\frac{N_2\left(\frac{h}{3}\right) - N_2(h)}{8}\right)$. Entonces $N_3(h)$ es una $O(h^3)$ aproximación a M.

11. Sea $N(h) = (1 + h)^{1/h}$, $N_2(h) = 2N\left(\frac{h}{2}\right) - N(h)$, $N_3(h) = N_2\left(\frac{h}{2}\right) + \frac{1}{3}\left(N_2\left(\frac{h}{2}\right) - N_2(h)\right)$.

a. $N(0.04) = 2.665836331$, $N(0.02) = 2.691588029$, $N(0.01) = 2.704813829$

b. $N_2(0.04) = 2.717339727$, $N_2(0.02) = 2.718039629$. La $O(h^3)$ aproximación es $N_3(0.04) = 2.718272931$.

c. Sí, porque los errores parecen proporcionales a h para $N(h)$, a h^2 para $N_2(h)$ y a h^3 para $N_3(h)$.

15. **c.**

k	4	8	16	32	64	128	256	512
p_k	$2\sqrt{2}$	3.0614675	3.1214452	3.1365485	3.1403312	3.1412723	3.1415138	3.1415729
P_k	4	3.3137085	3.1825979	3.1517249	3.144184	3.1422236	3.1417504	3.1416321

e. Los valores de p_k y P_k se incluyen en las tablas anexas, junto con los resultados de la extrapolación:

Para p_k

2.8284271				
3.0614675	3.1391476			
3.1214452	3.1414377	3.1415904		
3.1365485	3.1415829	3.1415926	3.1415927	
3.1403312	3.1415921	3.1415927	3.1415927	3.1415927

Para P_k

4				
3.3137085	3.0849447			
3.1825979	3.1388943	3.1424910		
3.1517249	3.1414339	3.1416032	3.1415891	
3.1441184	3.1415829	3.1415928	3.1415926	3.1415927

Conjunto de ejercicios 4.3

1. La regla del trapecio produce las siguientes aproximaciones. **a.** 0.265625 **b.** −0.2678571 **c.** −0.17776434 **d.** 0.1839397 **e.** −0.8666667 **f.** −0.1777643 **g.** 0.2180895 **h.** 4.1432597

3. La regla de Simpson produce las siguientes aproximaciones. **a.** 0.1940104 **b.** −0.2670635 **c.** 0.1922453 **d.** 0.16240168 **e.** −0.7391053 **f.** −0.1768216 **g.** 0.1513826 **h.** 2.5836964

5. La regla del punto medio produce las siguientes aproximaciones. **a.** 0.1582031 **b.** −0.2666667 **c.** 0.1743309 **d.** 0.1516327 **e.** −0.6753247 **f.** −0.1768200 **g.** 0.1180292 **h.** 1.8039148

7. $f(1) = \frac{1}{2}$

9. El grado de precisión es 3.

11. $c_0 = \frac{1}{3}, c_1 = \frac{4}{3}, c_2 = \frac{1}{3}$

13. $c_0 = c_1 = \frac{1}{2}$ produce el más alto grado de precisión, 1.

15. Las siguientes aproximaciones se obtienen de las fórmulas (4.23) a (4.30), respectivamente.

a. 0.1024404, 0.1024598, 0.1024598, 0.1024598, 0.1024695, 0.1024663, 0.1024598 y 0.1024598

b. 0.7853982, 0.7853982, 0.7853982, 0.7853982, 0.7853982, 0.7853982, 0.7853982 y 0.7853982

c. 1.497171, 1.477536, 1.477529, 1.477523, 1.467719, 1.470981, 1.477512 y 1.477515

d. 4.950000, 2.740909, 2.563393, 2.385700, 1.636364, 1.767857, 2.074893 y 2.116379

e. 3.293182, 2.407901, 2.359772, 2.314751, 1.965260, 2.048634, 2.233251 y 2.249001

f. 0.5000000, 0.6958004, 0.7126032, 0.7306341, 0.7937005, 0.7834709, 0.7611137 y 0.7593572

17. Los errores del ejercicio 16 son 1.6×10^{-6}, 5.3×10^{-8}, -6.7×10^{-7}, -7.2×10^{-7} y -1.3×10^{-6}, respectivamente.

19. Si $E(x^k) = 0$ para toda $k = 0, 1, \ldots, n$ y $E(x^{n+1}) \neq 0$, entonces con $p_{n+1}(x) = x^{n+1}$ tenemos un polinomio de grado $n + 1$ para el cual $E(p_{n+1}(x)) \neq 0$. Sea $p(x) = a_n x^n + \cdots + a_1 x + a_0$ cualquier polinomio de grado menor o igual que n. Entonces $E(p(x)) = a_n E(x^n) + \cdots + a_1 E(x) + a_0 E(1) = 0$. En cambio, si $E(p(x)) = 0$ para todos los polinomios de grado menor o igual que n, se deduce que $E(x^k) = 0$ para toda $k = 0, 1, \ldots, n$. Sea $p_{n+1}(x) = a_{n+1}x^{n+1} + \cdots + a_0$ un polinomio de grado $n + 1$ para el cual $E(p_{n+1}(x)) \neq 0$. Dado que $a_{n+1} \neq 0$, tenemos

$$x^{n+1} = \frac{1}{a_{n+1}} p_{n+1}(x) - \frac{a_n}{a_{n+1}} x^n - \cdots - \frac{a_0}{a_{n+1}}.$$

Entonces

$$E(x^{n+1}) = \frac{1}{a_{n+1}} E(P_{n+1}(x)) - \frac{a_n}{a_{n+1}} E(x_n) - \cdots - \frac{a_0}{a_{n+1}} E(1)$$
$$= \frac{1}{a_{n+1}} E(p_{n+1}(x)) \neq 0.$$

Por tanto, la fórmula de cuadratura tiene n grado de precisión.

Conjunto de ejercicios 4.4

1. Las aproximaciones obtenidas con la regla compuesta del trapecio son: **a.** 0.639900 **b.** 31.3653 **c.** 0.784241 **d.** −6.42872 **e.** −13.5760 **f.** 0.476977 **g.** 0.605498 **h.** 0.970926

3. Las aproximaciones obtenidas con la regla compuesta del punto medio son: **a.** 0.633096 **b.** 11.1568 **c.** 0.786700 **d.** −6.11274 **e.** −14.9985 **f.** 0.478751 **g.** 0.602961 **h.** 0.947868

5. $\alpha = 0.75$

7. a. La regla compuesta del trapecio requiere que $h < 0.000922295$ y que $n \geq 2168$.

b. La regla compuesta de Simpson requiere que $h < 0.037658$ y que $n \geq 54$.

c. La regla compuesta del punto medio requiere que $h < 0.00065216$ y que $n \geq 3066$.

9. a. La regla compuesta del trapecio requiere que $h < 0.04382$ y que $n \geq 46$. La aproximación es 0.405471.

b. La regla compuesta de Simpson requiere que $h < 0.44267$ y que $n \geq 6$. La aproximación es 0.405466.

c. La regla compuesta del punto medio requiere que $h < 0.03098$ y que $n \geq 64$. La aproximación es 0.405460.

11. a. Como los límites derecho e izquierdo en 0.1 y en 0.2 para f, f' y f'' son los mismos, las funciones serán continuas en [0, 0.3]. Sin embargo,

$$f'''(x) = \begin{cases} 6, & 0 \leq x \leq 0.1 \\ 12, & 0.1 < x \leq 0.2 \\ 12, & 0.2 < x \leq 0.3 \end{cases}$$

es discontinua en $x = 0.1$.

b. Tenemos 0.302506 con una cota de error de 1.9×10^{-4}.

c. Tenemos 0.302425 y el valor de la integral real es el mismo.

13. a. En el caso de la regla compuesta del trapecio, tenemos

$$E(f) = -\frac{h^3}{12}\sum_{j=1}^{n} f''(\xi_j) = -\frac{h^2}{12}\sum_{j=1}^{n} f''(\xi_j)h = -\frac{h^2}{12}\sum_{j=1}^{n} f''(\xi_j)\Delta x_j,$$

donde $\Delta x_j = x_{j+1} - x_j = h$ para toda j. Dado que $\sum_{j=1}^{n} f''(\xi_j)\,\Delta x_j$ es una suma de Riemann para $\int_a^b f''(x)dx = f'(b) - f'(a)$, tenemos

$$E(f) \approx -\frac{h^2}{12}[f'(b) - f'(a)].$$

b. En el caso de la regla compuesta del punto medio, tenemos

$$E(f) = -\frac{h^3}{3}\sum_{j=1}^{n/2} f''(\xi_j) = \frac{h^2}{6}\sum_{j=1}^{n/2} f''(\xi_j)(2h).$$

Pero $\sum_{j=1}^{n/2} f''(\xi_j)(2h)$ es una suma de Riemann para $\int_a^b f''(x)dx = f'(b) - f'(a)$, por lo que

$$E(f) \approx \frac{h^2}{6}[f'(b) - f'(a)].$$

15. a. La estimación mediante la regla compuesta del trapecio es $-\frac{1}{2}h^2 \ln 2 = -6.296 \times 10^{-6}$.

b. La estimación mediante la regla compuesta de Simpson es $-\frac{1}{240}h^2 = -3.75 \times 10^{-6}$.

c. La estimación mediante la regla del punto medio es $\frac{1}{6}h^2 \ln 2 = 6.932 \times 10^{-6}$.

17. La longitud es aproximadamente 15.8655.

19. La regla compuesta de Simpson con $h = 0.25$ da 2.61972 s.

21. La longitud es aproximadamente 58.47082 usando $n = 100$ en la regla compuesta de Simpson.

Conjunto de ejercicios 4.5

1. La integración de Romberg produce $R_{3,3}$ así: **a.** 0.1922593 **b.** 0.1606105 **c.** −0.1768200 **d.** 0.08875677 **e.** 2.5879685 **f.** −0.7341567 **g.** 0.6362135 **h.** 0.6426970

3. La integración de Romberg produce: **a.** 0.19225936 con $n = 4$ **b.** 0.16060279 con $n = 5$ **c.** −0.17682002 con $n = 4$ **d.** 0.088755284 con $n = 5$ **e.** 2.5886286 con $n = 6$ **f.** −0.73396918 con $n = 6$ **g.** 0.63621335 con $n = 4$ **h.** 0.64269908 con $n = 5$

5. $R_{33} = 11.5246$

7. $f(2.5) \approx 0.43459$

9. $R_{31} = 5$

11. Tenemos

$$R_{k,2} = \frac{4R_{k,1} - R_{k-1,1}}{3}$$

$$= \frac{1}{3}\left[R_{k-1,1} + 2h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (i - 1/2))h_{k-1})\right], \text{ de (4.35)},$$

$$= \frac{1}{3}\left[\frac{h_{k-1}}{2}(f(a) + f(b) + h_{k-1}\sum_{i=1}^{2^{k-2}-1} f(a + ih_{k-1})\right.$$

$$\left. + 2h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (i - 1/2)h_{k-1})\right], \text{ según (4.34) con } k - 1 \text{ en vez de } k,$$

$$= \frac{1}{3}\left[h_k(f(a) + f(b)) + 2h_k\sum_{i=1}^{2^{k-2}-1} f(a + 2ih_k) + 4h_k\sum_{i=1}^{2^{k-2}} f(a + (2i - 1)h)\right]$$

$$= \frac{h}{3}\left[f(a) + f(b) + 2\sum_{i=1}^{M-1} f(a + 2ih) + 4\sum_{i=1}^{M} f(a + (2i - 1)h)\right],$$

donde $h = h_k$ y $M = 2^{k-2}$.

13. La ecuación (4.35) proviene de

$$R_{k,1} = \frac{h_k}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{2^{k-1}-1} f(a + ih_k)\right]$$

$$= \frac{h_k}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{2^{k-1}-1} f(a + \frac{i}{2}h_{k-1})\right]$$

$$= \frac{h_k}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{2^{k-1}-1} f(a + ih_{k-1}) + 2\sum_{i=1}^{2^{k-2}} f(a + (i - 1/2)h_{k-1})\right]$$

$$= \frac{1}{2}\left\{\frac{h_{k-1}}{2}\left[f(a) + f(b) + 2\sum_{i=1}^{2^{k-2}-1} f(a + ih_{k-1})\right] + h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (i - 1/2)h_{k-1})\right\}$$

$$= \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (i - 1/2)h_{k-1})\right].$$

Conjunto de ejercicios 4.6

1. La regla de Simpson nos da

a. $S(1, 1.5) = 0.19224530$, $S(1, 1.25) = 0.039372434$, $S(1.25, 1.5) = 0.15288602$ y el valor real es 0.19225935.

b. $S(0, 1) = 0.16240168$, $S(0, 0.5) = 0.028861071$, $S(0.5, 1) = 0.13186140$ y el valor real es 0.16060279.

c. $S(0, 0.35) = -0.17682156$, $S(0, 0.175) = -0.087724382$, $S(0.175, 0.35) = -0.089095736$ y el valor real es -0.17682002.

d. $S(0, \frac{\pi}{4}) = 0.087995669$, $S(0, \frac{\pi}{8}) = 0.0058315797$, $S(\frac{\pi}{8}, \frac{\pi}{4}) = 0.082877624$ y el valor real es 0.088755285.

e. $S(0, \frac{\pi}{4}) = 2.5836964$, $S(0, \frac{\pi}{8}) = 0.33088926$, $S(\frac{\pi}{8}, \frac{\pi}{4}) = 2.2568121$ y el valor real es 2.5886286.

f. $S(1, 1.6) = -0.73910533$, $S(1, 1.3) = -0.26141244$, $S(1.3, 1.6) = -0.47305351$ y el valor real es -0.73396917.

g. $S(3, 3.5) = 0.63623873$, $S(3, 3.25) = 0.32567095$, $S(3.25, 3.5) = 0.31054412$ y el valor real es 0.63621334.

h. $S(0, \frac{\pi}{4}) = 0.64326905$, $S(0, \frac{\pi}{8}) = 0.37315002$, $S(\frac{\pi}{8}, \frac{\pi}{4}) = 0.26958270$ y el valor real es 0.64269908.

3. La cuadratura adaptiva produce: **a.** 108.555281 **b.** -1724.966983 **c.** -15.306308 **d.** -18.945949

5. La cuadratura adaptiva nos da

$$\int_{0.1}^{2} \operatorname{sen}\frac{1}{x}\,dx = 1.1454 \quad \text{y} \quad \int_{0.1}^{2} \cos\frac{1}{x}\,dx = 0.67378.$$

7. $\int_{0}^{2\pi} u(t)\,dt \approx 0.00001$

9.

t	$c(t)$	$s(t)$
0.1	0.0999975	0.000523589
0.2	0.199921	0.00418759
0.3	0.299399	0.0141166
0.4	0.397475	0.0333568
0.5	0.492327	0.0647203
0.6	0.581061	0.110498
0.7	0.659650	0.172129
0.8	0.722844	0.249325
0.9	0.764972	0.339747
1.0	0.779880	0.438245

Conjunto de ejercicios 4.7

1. La cuadratura gaussiana produce: **a.** 0.1922687 **b.** 0.1594104 **c.** -0.1768190 **d.** 0.08926302 **e.** 2.5913247 **f.** -0.7307230 **g.** 0.6361966 **h.** 0.6423172

3. La cuadratura gaussiana produce **a.** 0.1922594 **b.** 0.1606028 **c.** -0.1768200 **d.** 0.08875529 **e.** 2.5886327 **f.** -0.7339604 **g.** 0.6362133 **h.** 0.6426991

5. $a = 1$, $b = 1$, $c = \frac{1}{3}$, $d = -\frac{1}{3}$

Conjunto de ejercicios 4.8

1. El algoritmo 4.4 con $n = m = 4$ produce: **a.** 0.3115733 **b.** 0.2552526 **c.** 16.50864 **d.** 1.476684

3. El algoritmo 4.4 con $n = 4$ y con $m = 8$, $n = 8$ y $m = 4$, y $n = m = 6$ produce:

a. 0.5119875, 0.5118533, 0.5118722

b. 1.718857, 1.718220, 1.718385

c. 1.001953, 1.000122, 1.000386

d. 0.7838542, 0.7833659, 0.7834362

e. -1.985611, -1.999182, -1.997353

f. 2.004596, 2.000879, 2.000980

g. 0.3084277, 0.3084562, 0.3084323

h. -22.61612, -19.85408, -20.14117

5. El algoritmo 4.5 con $n = m = 2$ produce: **a.** 0.3115733 **b.** 0.2552446 **c.** 16.50863 **d.** 1.488875

7. El algoritmo 4.5 con $n = m = 3$, $n = 3$ y $m = 4$, $n = 4$ y $m = 3$, y $n = m = 4$ produce:

a. 0.5118655, 0.5118445, 0.5118655, 0.5118445, 2.1×10^{-5}, 1.3×10^{-7}, 2.1×10^{-5}, 1.3×10^{-7}

b. 1.718163, 1.718302, 1718139, 1.718277, 1.2×10^{-4}, 2.0×10^{-5}, 1.4×10^{-4}, 4.8×10^{-6}

c. 1.000000, 1.000000, 1.000000, 1.000000, 0, 0, 0, 0

d. 0.7833333, 0.7833333, 0.7833333, 0.7833333, 0, 0, 0, 0

e. −1.991878, −2.000124, −1.991878, −2.000124, 8.1×10^{-3}, 1.2×10^{-4}, 8.1×10^{-3}, 1.2×10^{-4}

f. 2.001494, 2.000080, 2.001388, 1.999984, 1.5×10^{-3}, 8×10^{-5}, 1.4×10^{-3}, 1.6×10^{-5}

g. 0.3084151, 0.3084145, 0.3084246, 0.3084245, 10^{-5}, 5.5×10^{-7}, 1.1×10^{-5}, 6.4×10^{-7}

h. −12.74790, −21.21539, −11.83624, −20.30373, 7.0, 1.5, 7.9, 0.564

9. El algoritmo 4.4 con $n = m = 14$ produce 0.1479103, y el algoritmo 4.5 con $n = m = 4$ produce 0.1506823.

11. La aproximación al centro de masa es $(\bar{x}, \bar{y})$, donde $\bar{x} = 0.3806333$ y $\bar{y} = 0.3822558$.

13. El área es aproximadamente 1.0402528.

15. El algoritmo 4.6 con $n = m = p = 2$ da el primer valor de la lista. El segundo es el resultado exacto.

a. 5.204036, $e(e^{0.5} - 1)(e - 1)^2$

b. 0.8429784, $\frac{1}{12}$

c. 0.08641975, $\frac{1}{14}$

d. 0.09722222, $\frac{1}{12}$

e. 7.103932, $2 + \frac{1}{12}\pi^2$

f. 1.428074, $\frac{1}{2}(e^2 + 1) - e$

17. El algoritmo 4.6 con $n = m = p = 4$ nos da el primer valor de la lista. El segundo proviene del algoritmo 4.6 con $n = m = p = 5$.

a. 5.206447, 5.206447

b. 0.08333333, 0.08333333

c. 0.07142857, 0.07142857

d. 0.08333333, 0.08333333

e. 6.934912, 6.934801

f. 1.476207, 1.476246

19. La aproximación 20.41887 requiere 125 evaluaciones funcionales.

Conjunto de ejercicios 4.9

1. La regla compuesta de Simpson produce: **a.** 0.5284163 **b.** 4.266654 **c.** 0.4329748 **d.** 0.8802210

3. La regla compuesta de Simpson produce: **a.** 04112649 **b.** 0.2440679 **c.** 0.05501681 **d.** 0.2903746

5. La velocidad de escape es aproximadamente 6.9450 mi/s.

7. a. $\int_0^\infty e^{-x} f(x)\,dx \approx 0.8535534\, f(0.5857864) + 0.1464466\, f(3.4142136)$

b. $\int_0^\infty e^{-x} f(x)\,dx \approx 0.7110930\, f(0.4157746) + 0.2785177\, f(2.2942804) + 0.0103893\, f(6.2899451)$

9. $n = 2$: 2.9865139

$n = 3$: 2.9958198

Conjunto de ejercicios 5.1

1. a. Dado que $f(t, y) = y \cos t$, tenemos $\frac{\partial f}{\partial y}(t, y) = \cos t$, y f satisface la condición de Lipschitz en y con $L = 1$ en

$$D = \{(t, y) \mid 0 \le t \le 1, -\infty < y < \infty\}$$

f es continua en D, así que existe una solución única. La solución es $y(t) = e^{\operatorname{sen} t}$.

b. $f(t, y) = \frac{2}{t} y + t^2 e^t$, $\frac{\partial f}{\partial y} = \frac{2}{t}$; satisface una condición de Lipschitz en y con $L = 2$ en

$$D = \{(t, y) \mid 1 \le t \le 2, -\infty < y < \infty\}$$

f es continua en D, así que existe una solución única, que es $y(t) = t^2(e^t - e)$.

c. $f(t, y) = -\frac{2}{t} y + t^2 e^t$, $\frac{\partial f}{\partial y} = -\frac{2}{t}$; f satisface una condición de Lipschitz en y con $L = 2$ en

$$D = \{(t, y) \mid 1 \le t \le 2, -\infty < y < \infty\}$$

f es continua en D, así que existe una solución única, que es

$$y(t) = (t^4 e^t - 4t^3 e^t + 12t^2 e^t - 24te^t + 24e^t + (\sqrt{2} - 9)e)/t^2.$$

d. $f(t, y) = \frac{4t^3 y}{1 + t^4}$, $\frac{\partial f}{\partial y} = \frac{4t^3}{1 + t^4}$ y f satisface una condición de Lipschitz en y con $L = 2$ en

$$D = \{(t, y) \mid 0 \le t \le 1, -\infty < y < \infty\}$$

f es continua en D, y, por tanto, existe una solución única, que es $y(t) = 1 + t^4$.

3. a. Al diferenciar $y^3 t + yt = 2$ se obtiene $3y^2 y' t + y^3 + y' t + y = 0$. Al resolver y' se obtiene la ecuación diferencial y, al utilizar $t = 1$ y $y = 1$, se verifica la condición inicial. Si quiere aproximar $y(2)$, aplique el método de Newton para resolver la ecuación $y^3 + y - 1 = 0$. Y así se obtiene $y(2) \approx 0.6823278$.

b. Al diferenciar $y \operatorname{sen} t + t^2 e^y + 2y - 1 = 0$ se obtiene $y' \operatorname{sen} t + y \cos t + 2te^y + t^2 e^y y' + 2y' = 0$. Al resolver y' se obtiene la ecuación diferencial original y, al usar $t = 1$ y $y = 0$, se verifica la condición inicial. Si quiere aproximar $y(2)$, aplique el método de Newton para resolver la ecuación $(2 + \operatorname{sen} 2)y + 4e^y - 1 = 0$. Y así obtendrá $y(2) \approx -0.4946599$.

5. Sean (t_1, y_1) y (t_2, y_2) en D, con $a \le t_1 \le b$, $a \le t_2 \le b$, $-\infty < y_1 < \infty$ y $-\infty < y_2 < \infty$. Para $0 \le \lambda \le 1$, tenemos $(1 - \lambda)a \le (1 - \lambda)t_1 \le (1 - \lambda)b$ por tanto, $\lambda a \le \lambda t_2 \le \lambda b$. Por otra parte, $a = (1 - \lambda)a + \lambda a \le (1 - \lambda)t_1 + \lambda t_2 \le (1 - \lambda)b + \lambda b = b$. Así que $-\infty < (1 - \lambda)y_1 + \lambda y_2 < \infty$, puesto que D es convexa.

7. a. Dado que $y' = f(t, y(t))$,

$$\int_a^t y'(z)\, dz = \int_a^t f(z, y(z))\, dz.$$

Por tanto,

$$y(t) - y(a) = \int_a^t f(z, y(z))\, dz$$

y

$$y(t) = \alpha + \int_a^t f(z, y(z))\, dz.$$

El método iterativo se deduce de esta ecuación.

b. Tenemos $y_0(t) = 1$, $y_1(t) = 1 + \frac{1}{2}t^2$, $y_2(t) = 1 + \frac{1}{2}t^2 - \frac{1}{6}t^3$, y $y_3(t) = 1 + \frac{1}{2}t^2 - \frac{1}{6}t^3 + \frac{1}{24}t^4$.

c. Tenemos $y(t) = 1 + \frac{1}{2}t^2 - \frac{1}{6}t^3 + \frac{1}{24}t^4 - \frac{1}{120}t^5 + \cdots$.

Conjunto de ejercicios 5.2

1. El método de Euler da las aproximaciones en las siguientes tablas.

a.

i	t_i	w_i	$y(t_i)$
1	0.500	0.0000000	0.2836165
2	1.000	1.1204223	3.2190993

b.

i	t_i	w_i	$y(t_i)$
1	2.500	2.0000000	1.8333333
2	3.000	2.6250000	2.5000000

c.

i	t_i	w_i	$y(t_i)$
1	1.250	2.7500000	2.7789294
2	1.500	3.5500000	3.6081977
3	1.750	4.3916667	4.4793276
4	2.000	5.2690476	5.3862944

d.

i	t_i	w_i	$y(t_i)$
1	0.250	1.2500000	1.3291498
2	0.500	1.6398053	1.7304898
3	0.750	2.0242547	2.0414720
4	1.000	2.2364573	2.1179795

3. El método de Euler da las aproximaciones en las siguientes tablas.

a.

i	t_i	w_i	$y(t_i)$
2	1.200	1.0082645	1.0149523
4	1.400	1.0385147	1.0475339
6	1.600	1.0784611	1.0884327
8	1.800	1.1232621	1.1336536
10	2.000	1.1706516	1.1812322

b.

i	t_i	w_i	$y(t_i)$
2	1.400	0.4388889	0.4896817
4	1.800	1.0520380	1.1994386
6	2.200	1.8842608	2.2135018
8	2.600	3.0028372	3.6784753
10	3.000	4.5142774	5.8741000

c.

i	t_i	w_i	$y(t_i)$
2	0.400	−1.6080000	−1.6200510
4	0.800	−1.3017370	−1.3359632
6	1.200	−1.1274909	−1.1663454
8	1.600	−1.0491191	−1.0783314
10	2.000	−1.0181518	−1.0359724

d.

i	t_i	w_i	$y(t_i)$
2	0.2	0.1083333	0.1626265
4	0.4	0.1620833	0.2051118
6	0.6	0.3455208	0.3765957
8	0.8	0.6213802	0.6461052
10	1.0	0.9803451	1.0022460

5. El método de Euler produce las aproximaciones que se incluyen en la tabla anexa.

a.

i	t_i	w_i	$y(t_i)$
1	1.1	0.271828	0.345920
5	1.5	3.18744	3.96767
6	1.6	4.62080	5.70296
9	1.9	11.7480	14.3231
10	2.0	15.3982	18.6831

b. La interpolación lineal produce las aproximaciones que se incluyen en la tabla anexa.

t	Aproximación	$y(t)$	Error
1.04	0.108731	0.119986	0.01126
1.55	3.90412	4.78864	0.8845
1.97	14.3031	17.2793	2.976

c. $h < 0.00064$

7. a. El método de Euler produce la siguiente aproximación a $y(5) = 5.00674$.

	$h = 0.2$	$h = 0.1$	$h = 0.05$
w_N	5.00377	5.00515	5.00592

b. $h = \sqrt{2 \times 10^{-6}} \approx 0.0014142$.

9. a. $h = 10^{-n/2}$ b. El error mínimo es $10^{-n/2}(e - 1) + 5e10^{-n-1}$.

c.

t	$w(h = 0.1)$	$w(h = 0.01)$	$y(t)$	Error ($n = 8$)
0.5	0.40951	0.39499	0.39347	1.5×10^{-4}
1.0	0.65132	0.63397	0.63212	3.1×10^{-4}

11. b. $w_{50} = 0.10430 \approx p(50)$ c. Dado que $p(t) = 1 - 0.99e^{-0.002t}$, $p(50) = 0.10421$.

Conjunto de ejercicios 5.3

1. a.

t_i	w_i	$y(t_i)$
0.50	0.12500000	0.28361652
1.00	2.02323897	3.21909932

b.

t_i	w_i	$y(t_i)$
2.50	1.75000000	1.83333333
3.00	2.42578125	2.50000000

c.

t_i	w_i	$y(t_i)$
1.25	2.78125000	2.77892944
1.50	3.61250000	3.60819766
1.75	4.48541667	4.47932763
2.00	5.39404762	5.38629436

d.

t_i	w_i	$y(t_i)$
0.25	1.34375000	1.32914981
0.50	1.77218707	1.73048976
0.75	2.11067606	2.04147203
1.00	2.20164395	2.11797955

3. a.

i	t_i	Orden 2 w_i	Orden 4 w_i	$y(t_i)$
1	1.1	1.214999	1.215883	1.215886
2	1.2	1.465250	1.467561	1.467570

b.

i	t_i	Orden 2	Orden 4 w_i	$y(t_i)$
1	0.5	0.5000000	0.5156250	0.5158868
2	1.0	1.076858	1.091267	1.091818

c.

i	t_i	Orden 2 w_i	Orden 4 w_i	$y(t_i)$
1	1.5	−2.000000	−2.000000	−1.500000
2	2.0	−1.777776	−1.679012	−1.333333
3	2.5	−1.585732	−1.484493	−1.250000
4	3.0	−1.458882	−1.374440	−1.200000

d.

i	t_i	Orden 2 w_i	Orden 4 w_i	$y(t_i)$
1	0.25	1.093750	1.086426	1.087088
2	0.50	1.312319	1.288245	1.289805
3	0.75	1.538468	1.512576	1.513490
4	1.0	1.720480	1.701494	1.701870

5. a. El método de segundo orden de Taylor produce los resultados que se incluyen en la tabla anexa.

i	t_i	w_i	$y(t_i)$
1	1.1	0.3397852	0.3459199
5	1.5	3.910985	3.967666
6	1.6	5.643081	5.720962
9	1.9	14.15268	14.32308
10	2.0	18.46999	18.68310

b. La interpolación lineal produce $y(1.04) \approx 0.1359139$, $y(1.55) \approx 4.777033$ y $y(1.97) \approx 17.17480$. Los valores reales son $y(1.04) = 0.1199875$, $y(1.55) = 4.788635$ y $y(1.97) = 17.27930$.

c. El método de cuarto orden de Taylor produce los resultados que se incluyen en la tabla anexa.

i	t_i	w_i
1	1.1	0.3459127
5	1.5	3.967603
6	1.6	5.720875
9	1.9	14.32290
10	2.0	18.68287

d. La interpolación cúbica de Hermite produce $y(1.04) \approx 0.1199704$, $y(1.55) \approx 4.788527$ y $y(1.97) \approx 17.27904$.

7. a.

i	t_i	Orden 2	Orden 4
2	0.2	5.86595	5.86433
5	0.5	2.82145	2.81789
7	0.7	0.84926	0.84455
10	1.0	−2.08606	−2.09015

b. 0.8 s

Conjunto de ejercicios 5.4

1. a.

t	Método modificado de Euler	$y(t)$
0.5	0.5602111	0.2836165
1.0	5.3014898	3.2190993

b.

t	Método modificado de Euler	$y(t)$
2.5	1.8125000	1.8333333
3.0	2.4815531	2.5000000

c.

t	Método modificado de Euler	$y(t)$
1.25	2.7750000	2.7789294
1.50	3.6008333	3.6081977
1.75	4.4688294	4.4793276
2.00	5.3728586	5.3862944

d.

t	Método modificado de Euler	$y(t)$
0.25	1.3199027	1.3291498
0.50	1.7070300	1.7304898
0.75	2.0053560	2.0414720
1.00	2.0770789	2.1179795

3. a.

t	Punto medio	$y(t)$
0.5	0.2646250	0.2836165
1.0	3.1300023	3.2190993

b.

t	Punto medio	$y(t)$
2.5	1.7812500	1.8333333
3.0	2.4550638	2.5000000

c.

t	Punto medio	$y(t)$
1.25	2.7777778	2.7789294
1.50	3.6060606	3.6081977
1.75	4.4763015	4.4793276
2.00	5.3824398	5.3862944

d.

t	Punto medio	$y(t)$
0.25	1.3337962	1.3291498
0.50	1.7422854	1.7304898
0.75	2.0596374	2.0414720
1.00	2.1385560	2.1179795

5. a. $1.0221167 \approx y(1.25) = 1.0219569$, $1.1640347 \approx y(1.93) = 1.1643901$
b. $1.9086500 \approx y(2.1) = 1.9249616$, $4.3105913 \approx y(2.75) = 4.3941697$
c. $-1.1461434 \approx y(1.3) = -1.1382768$, $-1.0454854 \approx y(1.93) = -1.0412665$
d. $0.3271470 \approx y(0.54) = 0.3140018$, $0.8967073 \approx y(0.94) = 0.8866318$

7. a. $1.0225530 \approx y(1.25) = 1.0219569$, $1.1646155 \approx y(1.93) = 1.1643901$
b. $1.9132167 \approx y(2.1) = 1.9249616$, $4.3246152 \approx y(2.75) = 4.3941697$
c. $-1.1441775 \approx y(1.3) = -1.1382768$, $-1.0447403 \approx y(1.93) = -1.0412665$
d. $0.3251049 \approx y(0.54) = 0.3140018$, $0.8945125 \approx y(0.94) = 0.8866318$

9. a. $1.0227863 \approx y(1.25) = 1.0219569$, $1.1649247 \approx y(1.93) = 1.1643901$
b. $1.9153749 \approx y(2.1) = 1.9249616$, $4.3312939 \approx y(2.75) = 4.3941697$
c. $-1.1432070 \approx y(1.3) = -1.1382768$, $-1.0443743 \approx y(1.93) = -1.0412665$
d. $0.3240839 \approx y(0.54) = 0.3140018$, $0.8934152 \approx y(0.94) = 0.8866318$

11. a. El método de cuarto orden de Runge-Kutta produce los resultados que se incluyen en las tablas anexas.

t	Runge-Kutta	$y(t)$
1.2	1.0149520	1.0149523
1.4	1.0475336	1.0475339
1.6	1.0884323	1.0884327
1.8	1.1336532	1.1336536
2.0	1.1812319	1.1812322

b.

t	Runge-Kutta	$y(t)$
1.4	0.4896842	0.4896817
1.8	1.1994320	1.1994386
2.2	2.2134693	2.2135018
2.6	3.6783790	3.6784753
3.0	5.8738386	5.8741000

c.

t	Runge-Kutta	$y(t)$
0.4	−1.6200576	−1.6200510
0.8	−1.3359824	−1.3359632
1.2	−1.1663735	−1.1663454
1.6	−1.0783582	−1.0783314
2.0	−1.0359922	−1.0359724

d.

t	Runge-Kutta	$y(t)$
0.2	0.1627655	0.1626265
0.4	0.2052405	0.2051118
0.6	0.3766981	0.3765957
0.8	0.6461896	0.6461052
0.2	1.0023207	1.0022460

15. En 0.2 tenemos aproximadamente 2099 unidades de KOH.

17. Las constantes apropiadas son $\alpha_1 = \delta_1 = \alpha_2 = \delta_2 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7 = \frac{1}{2}$ y $\alpha_3 = \delta_3 = 1$.

Conjunto de ejercicios 5.5

1. El algoritmo de Runge-Kutta-Fehlberg produce los resultados que se incluyen en las tablas anexas.

a.

i	t_i	w_i	h_i	y_i
1	0.2093900	0.0298184	0.2093900	0.0298337
3	0.5610469	0.4016438	0.1777496	0.4016860
5	0.8387744	1.5894061	0.1280905	1.5894600
7	1.0000000	3.2190497	0.0486737	3.2190993

b.

i	t_i	w_i	h_i	y_i
1	2.2500000	1.4499988	0.2500000	1.4500000
2	2.5000000	1.8333332	0.2500000	1.8333333
3	2.7500000	2.1785718	0.2500000	2.1785714
4	3.0000000	2.5000005	0.2500000	2.5000000

c.

i	t_i	w_i	h_i	y_i
1	1.2500000	2.7789299	0.2500000	2.7789294
2	1.5000000	3.6081985	0.2500000	3.6081977
3	1.7500000	4.4793288	0.2500000	4.4793276
4	2.0000000	5.3862958	0.2500000	5.3862944

d.

i	t_i	w_i	h_i	y_i
1	0.2500000	1.3291478	0.2500000	1.3291498
2	0.5000000	1.7304857	0.2500000	1.7304898
3	0.7500000	2.0414669	0.2500000	2.0414720
4	1.0000000	2.1179750	0.2500000	2.1179795

3. El algoritmo de Runge-Kutta-Fehlberg produce los resultados que se incluyen en las tablas anexas.

a.

i	t_i	w_i	h_i	y_i
1	1.1101946	1.0051237	0.1101946	1.0051237
5	1.7470584	1.1213948	0.2180472	1.1213947
7	2.3994350	1.2795396	0.3707934	1.2795395
11	4.0000000	1.6762393	0.1014853	1.6762391

b.

i	t_i	w_i	h_i	y_i
4	1.5482238	0.7234123	0.1256486	0.7234119
7	1.8847226	1.3851234	0.1073571	1.3851226
10	2.1846024	2.1673514	0.0965027	2.1673499
16	2.6972462	4.1297939	0.0778628	4.1297904
21	3.0000000	5.8741059	0.0195070	5.8741000

c.

i	t_i	w_i	h_i	y_i
1	0.1633541	−1.8380836	0.1633541	−1.8380836
5	0.7585763	−1.3597623	0.1266248	−1.3597624
9	1.1930325	−1.1684827	0.1048224	−1.1684830
13	1.6229351	−1.0749509	0.1107510	−1.0749511
17	2.1074733	−1.0291158	0.1288897	−1.0291161
23	3.0000000	−1.0049450	0.1264618	−1.0049452

d.

i	t_i	w_i	h_i	y_i
1	0.3986051	0.3108201	0.3986051	0.3108199
3	0.9703970	0.2221189	0.2866710	0.2221186
5	1.5672905	0.1133085	0.3042087	0.1133082
8	2.0000000	0.0543454	0.0902302	0.0543455

5. a. El número de infectados es $y(30) = 80295.7$

b. En este modelo, el valor límite del número de infectados es $\lim_{t\to\infty} y(t) = 100000$.

Conjunto de ejercicios 5.6

1. Los métodos de Adams-Bashforth producen los resultados que se incluyen en las tablas anexas.

a.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
0.2	0.0268128	0.0268128	0.0268128	0.0268128	0.0268128
0.4	0.1200522	0.1507778	0.1507778	0.1507778	0.1507778
0.6	0.4153551	0.4613866	0.4960196	0.4960196	0.4960196
0.8	1.1462844	1.2512447	1.2961260	1.3308570	1.3308570
1.0	2.8241683	3.0360680	3.1461400	3.1854002	3.2190993

b.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
2.2	1.3666667	1.3666667	1.3666667	1.3666667	1.3666667
2.4	1.6750000	1.6857143	1.6857143	1.6857143	1.6857143
2.6	1.9632431	1.9794407	1.9750000	1.9750000	1.9750000
2.8	2.2323184	2.2488759	2.2423065	2.2444444	2.2444444
3.0	2.4884512	2.5051340	2.4980306	2.5011406	2.5000000

c.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
1.2	2.6187859	2.6187859	2.6187859	2.6187859	2.6187859
1.4	3.2734823	3.2710611	3.2710611	3.2710611	3.2710611
1.6	3.9567107	3.9514231	3.9520058	3.9520058	3.9520058
1.8	4.6647738	4.6569191	4.6582078	4.6580160	4.6580160
2.0	5.3949416	5.3848058	5.3866452	5.3862177	5.3862944

d.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
0.2	1.2529306	1.2529306	1.2529306	1.2529306	1.2529306
0.4	1.5986417	1.5712255	1.5712255	1.5712255	1.5712255
0.6	1.9386951	1.8827238	1.8750869	1.8750869	1.8750869
0.8	2.1766821	2.0844122	2.0698063	2.0789180	2.0789180
1.0	2.2369407	2.1115540	2.0998117	2.1180642	2.1179795

3. Los métodos de Adams-Bashforth producen los resultados que se incluyen en las tablas anexas.

a.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
1.2	1.0161982	1.0149520	1.0149520	1.0149520	1.0149523
1.4	1.0497665	1.0468730	1.0477278	1.0475336	1.0475339
1.6	1.0910204	1.0875837	1.0887567	1.0883045	1.0884327
1.8	1.1363845	1.1327465	1.1340093	1.1334967	1.1336536
2.0	1.1840272	1.1803057	1.1815967	1.1810689	1.1812322

b.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
1.4	0.4867550	0.4896842	0.4896842	0.4896842	0.4896817
1.8	1.1856931	1.1982110	1.1990422	1.1994320	1.1994386
2.2	2.1753785	2.2079987	2.2117448	2.2134792	2.2135018
2.6	3.5849181	3.6617484	3.6733266	3.6777236	3.6784753
3.0	5.6491203	5.8268008	5.8589944	5.8706101	5.8741000

c.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
0.5	−1.5357010	−1.5381988	−1.5379372	−1.5378676	−1.5378828
1.0	−1.2374093	−1.2389605	−1.2383734	−1.2383693	−1.2384058
1.5	−1.0952910	−1.0950952	−1.0947925	−1.0948481	−1.0948517
2.0	−1.0366643	−1.0359996	−1.0359497	−1.0359760	−1.0359724

d.

t	Paso 2	Paso3	Paso 4	Paso 5	$y(t)$
0.2	0.1739041	0.1627655	0.1627655	0.1627655	0.1626265
0.4	0.2144877	0.2026399	0.2066057	0.2052405	0.2051118
0.6	0.3822803	0.3747011	0.3787680	0.3765206	0.3765957
0.8	0.6491272	0.6452640	0.6487176	0.6471458	0.6461052
1.0	1.0037415	1.0020894	1.0064121	1.0073348	1.0022460

5. Los métodos de predictor-corrector de cuarto orden de Adams producen los resultados que se incluyen en las tablas anexas.

a.

t	w	$y(t)$
1.2	1.0149520	1.0149523
1.4	1.0475227	1.0475339
1.6	1.0884141	1.0884327
1.8	1.1336331	1.1336536
2.0	1.1812112	1.1812322

b.

t	w	$y(t)$
1.4	0.4896842	0.4896817
1.8	1.1994245	1.1994386
2.2	2.2134701	2.2135018
2.6	3.6784144	3.6784753
3.0	5.8739518	5.8741000

c.

t	w	$y(t)$
0.5	−1.5378788	−1.5378828
1.0	−1.2384134	−1.2384058
1.5	−1.0948609	−1.0948517
2.0	−1.0359757	−1.0359724

d.

t	w	$y(t)$
0.2	0.1627655	0.1626265
0.4	0.2048557	0.2051118
0.6	0.3762804	0.3765957
0.8	0.6458949	0.6461052
1.0	1.0021372	1.0022460

7. a. Con $h = 0.01$, el método de tres pasos de Adams-Moulton produce los valores de la tabla anexa.

i	t_i	w_i
10	0.1	1.317218
20	0.2	1.784511

b. El método de Newton reducirá de tres a dos el número de iteraciones por paso, aplicando el criterio de parar o detención

$$\left| w_i^{(k)} - w_i^{(k-1)} \right| \leq 10^{-6}.$$

13. Para derivar el método de Milne, se integra $y'(t) = f(t, y(t))$ en el intervalo $[t_{i-3}, t_{i+1}]$ para obtener

$$y(t_{i+1}) - y(t_{i-3}) = \int_{t_{i-3}}^{t_{i+1}} f(t, y(t))\, dt.$$

Al usar la fórmula abierta de Newton-Cotes (4.29), tenemos

$$y(t_{i+1}) - y(t_{i-3}) = \frac{4h[2f(t_i, y(t_i)) - f(t_{i-1}, y(t_{i-1})) + 2f(t_{i-2}, y(t_{i-2}))]}{} + \frac{14h^5 f^{(4)}(\xi, y(\xi))}{45}.$$

La ecuación de diferencias se convierte en

$$w_{i+1} = w_{i-3} + \frac{h[8f(t_i, w_i) - 4f(t_{i-1}, w_{i-1}) + 8f(t_{i-2}, w_{i-2})]}{},$$

con el error local de truncamiento

$$\tau_{i+1}(h) = \frac{14h^4 y^{(5)}(\xi)}{45}.$$

Conjunto de ejercicios 5.7

1. El algoritmo predictor-corrector de tamaño de paso variable de Adams produce los resultados que se incluyen en las tablas anexas.

a.

i	t_i	w_i	h_i	y_i
1	0.04275596	0.00096891	0.04275596	0.00096887
5	0.22491460	0.03529441	0.05389076	0.03529359
12	0.60214994	0.50174348	0.05389076	0.50171761
17	0.81943926	1.45544317	0.04345786	1.45541453
22	0.99830392	3.19605697	0.03577293	3.19602842
26	1.00000000	3.21912776	0.00042395	3.21909932

b.

i	t_i	w_i	h_i	y_i
1	2.06250000	1.12132350	0.06250000	1.12132353
5	2.31250000	1.55059834	0.06250000	1.55059524
9	2.62471924	2.00923157	0.09360962	2.00922829
13	2.99915773	2.49895243	0.09360962	2.49894707
17	3.00000000	2.50000535	0.00021057	2.50000000

c.

i	t_i	w_i	h_i	y_i
1	1.06250000	2.18941363	0.06250000	2.18941366
4	1.25000000	2.77892931	0.06250000	2.77892944
8	1.85102559	4.84179835	0.15025640	4.84180141
12	2.00000000	5.38629105	0.03724360	5.38629436

d.

i	t_i	w_i	h_i	y_i
1	0.06250000	1.06817960	0.06250000	1.06817960
5	0.31250000	1.42861668	0.06250000	1.42861361
10	0.62500000	1.90768386	0.06250000	1.90767015
13	0.81250000	2.08668486	0.06250000	2.08666541
16	1.00000000	2.11800208	0.06250000	2.11797955

3. Las tablas anexas contienen resultados representativos obtenidos con el algoritmo predictor-corrector de tamaño variable de paso de Adams.

a.

i	t_i	w_i	h_i	y_i
5	1.10431651	1.00463041	0.02086330	1.00463045
15	1.31294952	1.03196889	0.02086330	1.03196898
25	1.59408142	1.08714711	0.03122028	1.08714722
35	2.00846205	1.18327922	0.04824992	1.18327937
45	2.66272188	1.34525123	0.07278716	1.34525143
52	3.40193112	1.52940900	0.11107035	1.52940924
57	4.00000000	1.67623887	0.12174963	1.67623914

b.

i	t_i	w_i	h_i	y_i
5	1.18519603	0.20333499	0.03703921	0.20333497
15	1.55558810	0.73586642	0.03703921	0.73586631
25	1.92598016	1.48072467	0.03703921	1.48072442
35	2.29637222	2.51764797	0.03703921	2.51764743
45	2.65452689	3.92602442	0.03092051	3.92602332
55	2.94341188	5.50206466	0.02584049	5.50206279
61	3.00000000	5.87410206	0.00122679	5.87409998

c.

i	t_i	w_i	h_i	y_i
5	0.16854008	−1.83303780	0.03370802	−1.83303783
17	0.64833341	−1.42945306	0.05253230	−1.42945304
27	1.06742915	−1.21150951	0.04190957	−1.21150932
41	1.75380240	−1.05819340	0.06681937	−1.05819325
51	2.50124702	−1.01335240	0.07474446	−1.01335258
61	3.00000000	−1.00494507	0.01257155	−1.00494525

d.

i	t_i	w_i	h_i	y_i
5	0.28548652	0.32153668	0.05709730	0.32153674
15	0.85645955	0.24281066	0.05709730	0.24281095
20	1.35101725	0.15096743	0.09891154	0.15096772
25	1.66282314	0.09815109	0.06236118	0.09815137
29	1.91226786	0.06418555	0.06236118	0.06418579
33	2.00000000	0.05434530	0.02193303	0.05434551

5. La corriente después de 2 s es aproximadamente $i(2) = 8.693$ amperes.

Conjunto de ejercicios 5.8

1. El algoritmo de extrapolación produce los resultados que se incluyen en las tablas anexas.

a.

i	t_i	w_i	h	k	y_i
1	0.25	0.04543132	0.25	3	0.04543123
2	0.50	0.28361684	0.25	3	0.28361652
3	0.75	1.05257634	0.25	4	1.05257615
4	1.00	3.21909944	0.25	4	3.21909932

b.

i	t_i	w_i	h	k	y_i
1	2.25	1.44999987	0.25	3	1.45000000
2	2.50	1.83333321	0.25	3	1.83333333
3	2.75	2.17857133	0.25	3	2.17857143
4	3.00	2.49999993	0.25	3	2.50000000

c.

i	t_i	w_i	h	k	y_i
1	1.25	2.77892942	0.25	3	2.77892944
2	1.50	3.60819763	0.25	3	3.60819766
3	1.75	4.47932759	0.25	3	4.47932763
4	2.00	5.38629431	0.25	3	5.38629436

d.

i	t_i	w_i	h	k	y_i
1	0.25	1.32914981	0.25	3	1.32914981
2	0.50	1.73048976	0.25	3	1.73048976
3	0.75	2.04147203	0.25	3	2.04147203
4	1.00	2.11797954	0.25	3	2.11797955

3. El algoritmo de extrapolación produce los resultados que se incluyen en las tablas anexas.

a.

i	t_i	w_i	h	k	y_i
1	1.50	1.06726237	0.50	4	1.06726235
2	2.00	1.18123223	0.50	3	1.18123222
3	2.50	1.30460372	0.50	3	1.30460371
4	3.00	1.42951608	0.50	3	1.42951607
5	3.50	1.55364771	0.50	3	1.55364770
6	4.00	1.67623915	0.50	3	1.67623914

b.

i	t_i	w_i	h	k	y_i
1	1.50	0.64387537	0.50	4	0.64387533
2	2.00	1.66128182	0.50	5	1.66128176
3	2.50	3.25801550	0.50	5	3.25801536
4	3.00	5.87410027	0.50	5	5.87409998

c.

i	t_i	w_i	h	k	y_i
1	0.50	−1.53788284	0.50	4	−1.53788284
2	1.00	−1.23840584	0.50	5	−1.23840584
3	1.50	−1.09485175	0.50	5	−1.09485175
4	2.00	−1.03597242	0.50	5	−1.03597242
5	2.50	−1.01338570	0.50	5	−1.01338570
6	3.00	−1.00494526	0.50	4	−1.00494525

d.

i	t_i	w_i	h	k	y_i
1	0.50	0.29875177	0.50	4	0.29875178
2	1.00	0.21662642	0.50	4	0.21662642
3	1.50	0.12458565	0.50	4	0.12458565
4	2.00	0.05434552	0.50	4	0.05434551

Conjunto de ejercicios 5.9

1. El algoritmo de Runge-Kutta para sistemas produce los resultados que se incluyen en las tablas anexas.

a.

t_i	w_{1i}	u_{1i}	w_{2i}	u_{2i}
0.200	2.12036583	2.12500839	1.50699185	1.51158743
0.400	4.44122776	4.46511961	3.24224021	3.26598528
0.600	9.73913329	9.83235869	8.16341700	8.25629549
0.800	22.67655977	23.00263945	21.34352778	21.66887674
1.000	55.66118088	56.73748265	56.03050296	57.10536209

b.

t_i	w_{1i}	u_{1i}	w_{2i}	u_{2i}
0.500	0.95671390	0.95672798	−1.08381950	−1.08383310
1.000	1.30654440	1.30655930	−0.83295364	−0.83296776
1.500	1.34416716	1.34418117	−0.56980329	−0.56981634
2.000	1.14332436	1.14333672	−0.36936318	−0.36937457

c.

t_i	w_{1i}	u_{1i}	w_{2i}	u_{2i}	w_{3i}	u_{3i}
0.5	0.70787076	0.70828683	−1.24988663	−1.25056425	0.39884862	0.39815702
1.0	−0.33691753	−0.33650854	−3.01764179	−3.01945051	−0.29932294	−0.30116868
1.5	−2.41332734	−2.41345688	−5.40523279	−5.40844686	−0.92346873	−0.92675778
2.0	−5.89479008	−5.89590551	−8.70970537	−8.71450036	−1.32051165	−1.32544426

d.

t_i	w_{1i}	u_{1i}	w_{2i}	u_{2i}	w_{3i}	u_{3i}
0.2	1.38165297	1.38165325	1.00800000	1.00800000	−0.61833075	−0.61833075
0.5	1.90753116	1.90753184	1.12500000	1.12500000	−0.09090565	−0.09090566
0.7	2.25503524	2.25503620	1.34300000	1.34300000	0.26343971	0.26343970
1.0	2.83211921	2.83212056	2.00000000	2.00000000	0.88212058	0.88212056

5. El método predictor-corrector de cuarto orden de Adams para sistemas produce los resultados que se incluyen en las tablas anexas.

a.

t_i	w_{1i}	$y(t_i)$
0.200	0.00015352	0.00015350
0.500	0.00743133	0.00743027
0.700	0.03300266	0.03299805
1.000	0.17134711	0.17132880

b.

t_i	w_{1i}	$y(t_i)$
1.200	0.96152437	0.96152583
1.500	0.77796798	0.77797237
1.700	0.59373213	0.59373830
2.000	0.27258055	0.27258872

c.

t_i	w_{1i}	$y(t_i)$
1.000	3.73186337	3.73170445
2.000	11.31462595	11.31452924
3.000	34.04548233	34.04517155

d.

t_i	w_{1i}	$y(t_i)$
1.200	0.27273759	0.27273791
1.500	1.08847933	1.08849259
1.700	2.04352376	2.04353642
2.000	4.36157310	4.36157780

7. Los números predichos de presas, x_{1i}, y de depredadores, x_{2i}, se incluyen en la tabla anexa.

i	t_i	x_{1i}	x_{2i}
10	1.0	4393	1512
20	2.0	288	3175
30	3.0	32	2042
40	4.0	25	1258

Conjunto de ejercicios 5.10

1. Sea L la constante de Lipschitz para ϕ. Entonces

$$u_{i+1} - v_{i+1} = u_i - v_i + h[\phi(t_i, u_i, h) - \phi(t_i, v_i, h)],$$

así que

$$|u_{i+1} - v_{i+1}| \leq (1 + hL)|u_i - v_i| \leq (1 + hL)^{i+1}|u_0 - v_0|.$$

3. De acuerdo con el ejercicio 17 de la sección 5.4, tenemos

$$\begin{aligned}\phi(t, w, h) = {} & \frac{1}{6}f(t, w) + \frac{1}{3}f\left(t + \frac{1}{2}h, w + \frac{1}{2}hf(t, w)\right) \\ & + \frac{1}{3}f\left(t + \frac{1}{2}h, w + \frac{1}{2}hf\left(t + \frac{1}{2}h, w + \frac{1}{2}hf(t, w)\right)\right) \\ & + \frac{1}{6}f\left(t + h, w + hf\left(t + \frac{1}{2}h, w + \frac{1}{2}hf\left(t + \frac{1}{2}h, w + \frac{1}{2}hf(t, w)\right)\right)\right).\end{aligned}$$

y, por tanto,

$$\phi(t, w, 0) = \frac{1}{6}f(t, w) + \frac{1}{3}f(t, w) + \frac{1}{3}f(t, w) + \frac{1}{6}f(t, w) = f(t, w).$$

5. a. El error local de truncamiento es $\tau_{i+1} = \frac{1}{4}h^3y^{(4)}(\xi_i)$, para alguna ξ, donde $t_{i-2} < \xi_i < t_{i+1}$.

b. El método es consistente pero inestable y no convergente.

7. El método es inestable.

Conjunto de ejercicios 5.11

1. El método de Euler produce los resultados que se incluyen en las tablas anexas.

a.

t_i	w_i	y_i
0.200	0.027182818	0.449328964
0.500	0.000027183	0.030197383
0.700	0.000000272	0.004991594
1.000	0.000000000	0.000335463

b.

t_i	w_i	y_i
0.200	0.373333333	0.046105213
0.500	−0.933333333	0.250015133
0.700	0.146666667	0.490000277
1.000	1.333333333	1.000000001

c.

t_i	w_i	y_i
0.500	16.47925	0.479470939
1.000	256.7930	0.841470987
1.500	4096.142	0.997494987
2.000	65523.12	0.909297427

d.

t_i	w_i	y_i
0.200	6.128259	1.000000001
0.500	−378.2574	1.000000000
0.700	6052.063	1.000000000
1.000	387332.0	1.000000000

3. El algoritmo predictor-corrector de cuarto orden de Adams produce los resultados que se incluyen en las tablas anexas.

a.

t_i	w_i	y_i
0.200	0.4588119	0.4493290
0.500	−0.0112813	0.0301974
0.700	0.0013734	0.0049916
1.000	0.0023604	0.0003355

b.

t_i	w_i	y_i
0.200	0.0792593	0.0461052
0.500	0.1554027	0.2500151
0.700	0.5507445	0.4900003
1.000	0.7278557	1.0000000

c.

t_i	w_i	y_i
.500	188.3082	0.4794709
1.000	38932.03	0.8414710
1.500	9073607	0.9974950
2.000	2115741299	0.9092974

d.

t_i	w_i	y_i
0.200	−215.7459	1.000000001
0.500	−682637.0	1.000000000
0.700	−159172736	1.000000000
1.000	−566751172258	1.000000000

5. a.

t_i	w_{1i}	u_{1i}	w_{2i}	u_{2i}
0.100	−96.33011	0.66987648	193.6651	−0.33491554
0.200	−28226.32	0.67915383	56453.66	−0.33957692
0.300	−8214056	0.69387881	16428113	−0.34693941
0.400	−2390290586	0.71354670	4780581173	−0.35677335
0.500	−695574560790	0.73768711	1391149121600	−0.36884355

b.

t_i	w_{1i}	u_{1i}	w_{2i}	u_{2i}
0.100	0.61095960	0.66987648	−0.21708179	−0.33491554
0.200	0.66873489	0.67915383	−0.31873903	−0.33957692
0.300	0.69203679	0.69387881	−0.34325535	−0.34693941
0..400	0.71322103	0.71354670	−0.35612202	−0.35677335
0.500	0.73762953	0.73768711	−0.36872840	−0.36884355

9. a. El método del trapecio aplicado a la ecuación de prueba produce

$$w_{j+1} = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}}\, w_j,$$

así que

$$Q(h\lambda) = \frac{2 + h\lambda}{2 - h\lambda}.$$

Por tanto, $|Q(h\lambda)| < 1$, siempre que $\text{Re}(h\lambda) < 0$.

b. El método regresivo de Euler aplicado a la ecuación de prueba produce

$$w_{j+1} = \frac{w_j}{1 - h\lambda},$$

así que

$$Q(h\lambda) = \frac{1}{1 - h\lambda}.$$

Por tanto, $|Q(h\lambda)| < 1$, siempre que $\text{Re}(h\lambda) < 0$.

Conjunto de ejercicios 6.1

1. a. Líneas intersecantes con la solución $x_1 = x_2 = 1$.

b. Líneas intersecantes con la solución $x_1 = x_2 = 0$.

c. Una línea y, por tanto, hay una cantidad infinita de soluciones $x_2 = \frac{3}{2} - \frac{1}{2}x_1$.

d. Líneas paralelas y, por tanto, no hay una solución.

e. Una línea y, por tanto, hay una cantidad infinita de soluciones con $x_2 = -\frac{1}{2}x_1$.

f. Tres líneas en el plano que no se intersecan en un punto común.

g. Las líneas intersecantes con la solución $x_1 = \frac{2}{7}$ y $x_2 = -\frac{11}{7}$.

h. Dos planos en el espacio que se intersecan en una línea con $x_1 = -\frac{5}{4}x_2$ y $x_3 = \frac{3}{2}x_2 + 1$.

3. La eliminación gaussiana produce las siguientes soluciones.

a. $x_1 = 1.1875$, $x_2 = 1.8125$, $x_3 = 0.875$, requiriéndose un intercambio de un renglón.

b. $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, sin que se requiera intercambio alguno.

c. $x_1 = 1.5$, $x_2 = 2$, $x_3 = -1.2$, $x_4 = 3$ sin que se requiera intercambio alguno.

d. $x_1 = \frac{22}{9}$, $x_2 = -\frac{4}{9}$, $x_3 = \frac{4}{3}$, $x_4 = 1$, requiriéndose un intercambio de un renglón.

e. No existe una solución única.

f. $x_1 = -1$, $x_2 = 2$, $x_3 = 0$, $x_4 = 1$, requiriéndose un intercambio de un renglón.

5. a. Cuando $\alpha = -1/3$, no existe solución.

b. Cuando $\alpha = 1/3$, existe una cantidad infinita de soluciones con $x_1 = x_2 + 1.5$ y x_2 es arbitraria.

c. Si $\alpha \neq \pm 1/3$, entonces la solución única será

$$x_1 = \frac{3}{2(1 + 3\alpha)} \text{ y } x^2 = \frac{-3}{2(1 + 3\alpha)}.$$

9. El método de Gauss-Jordan produce los siguientes resultados.

a. $x_1 = 0.98$, $x_2 = -0.98$, $x_3 = 2.9$

b. $x_1 = 1.1$, $x_2 = -1.0$, $x_3 = 2.9$

11. b. Los resultados de este ejercicio se dan en la tabla anexa. (Con las abreviaturas M/D y A/S designamos multiplicaciones/divisiones y adiciones/sustracciones, respectivamente.)

	Eliminación gaussiana		Método de Gauss-Jordan	
n	M/D	A/S	M/D	A/S
3	17	11	21	12
10	430	375	595	495
50	44150	42875	64975	62475
100	343300	338250	509950	499950

13. El método híbrido de eliminación gaussiana y Gauss-Jordan produce los siguientes resultados.

a. $x_1 = 1.0$, $x_2 = -0.98$, $x_3 = 2.9$

b. $x_1 = 1.0$, $x_2 = -1.0$, $x_3 = 2.9$

15. a. Hay suficiente comida para satisfacer el consumo diario promedio.

b. Podríamos agregar 200 de la especie 1, 150 de la especie 2, 100 de la especie 3 o 100 de la especie 4.

c. Suponiendo que no se haya seleccionado ninguno de los aumentos indicados en el inciso (b), la especie 2 podría incrementarse en 650, la especie 3 en 150 o la especie 4 en 150.

d. Suponiendo que no se haya seleccionado ninguno de los incrementos indicados en los incisos (b) o (c), la especie 3 podría aumentarse en 150 o la especie 4 en 150.

Conjunto de ejercicios 6.2

1. a. Ninguno. **b.** Intercambiar renglones 2 y 3. **c.** Ninguno. **d.** Intercambiar renglones 1 y 2.

3. a. Intercambiar renglones 1 y 3 y luego intercambiar renglones 2 y 3.

b. Intercambiar renglones 2 y 3.

c. Intercambiar renglones 2 y 3.

d. Intercambiar renglones 1 y 3 y luego los renglones 2 y 3.

5. La eliminación gaussiana con la aritmética de corte a tres dígitos produce los siguientes resultados.

a. $x_1 = 30.0$, $x_2 = 0.990$

b. $x_1 = 1.00$, $x_2 = 9.98$

c. $x_1 = 0.00$, $x_2 = 10.0$, $x_3 = 0.142$

d. $x_1 = 12.0$, $x_2 = 0.492$, $x_3 = -9.78$

e. $x_1 = 0.206$, $x_2 = 0.0154$, $x_3 = -0.0156$, $x_4 = -0.716$

f. $x_1 = 0.828$, $x_2 = -3.32$, $x_3 = 0.153$, $x_4 = 4.91$

7. La eliminación gaussiana con pivoteo parcial y con la aritmética de corte a tres dígitos produce los siguientes resultados.

a. $x_1 = 10.0$, $x_2 = 1.00$

b. $x_1 = 1.00$, $x_2 = 9.98$

c. $x_1 = -0.163$, $x_2 = 9.98$, $x_3 = 0.142$

d. $x_1 = 12.0, x_2 = 0.504, x_3 = -9.78$
e. $x_1 = 0.177, x_2 = 0.0072, x_3 = -0.0208, x_4 = -1.18$
f. $x_1 = 0.777, x_2 = -3.10, x_3 = 0.161, x_4 = 4.50$

9. La eliminación gaussiana con pivoteo parcial escalado y con aritmética de corte a tres dígitos produce los siguientes resultados.
a. $x_1 = 10.0, x_2 = 1.00$
b. $x_1 = 1.00, x_2 = 9.98$
c. $x_1 = -0.163, x_2 = 9.98, x_3 = 0.142$
d. $x_1 = 0.993, x_2 = 0.500, x_3 = -1.00$
e. $x_1 = 0.171, x_2 = 0.0102, x_3 = -0.0217, x_4 = -1.27$
f. $x_1 = 0.687, x_2 = -2.66, x_3 = 0.117, x_4 = 3.59$

11. La eliminación gaussiana con el algoritmo de sustitución regresiva y la aritmética de precisión simple produce los siguientes resultados.
Para (1a) tenemos $x_1 = 10.000000, x_2 = 1.0000000$.
Para (1b) tenemos $x_1 = 1.0000000, x_2 = 10.000000$.
Para (1c) tenemos $x_1 = 0.0000000, x_2 = 10.000000, x_3 = 0.14285714$.
Para (1d) tenemos $x_1 = 0.99999999, x_2 = 0.50000000, x_3 = -1.00000000$.
Para (1e) tenemos $x_1 = 0.17682530, x_2 = 0.012692691, x_3 = -0.020654050, x_4 = -1.1826087$.
Para (1f) tenemos $x_1 = 0.78838790, x_2 = -3.1253894, x_3 = 0.1675964, x_4 = 4.5569519$.

13. La eliminación gaussiana con el algoritmo de pivoteo parcial escalado y con la aritmética de precisión simple produce los siguientes resultados.
Para (1a) tenemos $x_1 = 10.000000, x_2 = 1.0000000$.
Para (1b) tenemos $x_1 = 1.0000000, x_2 = 10.000000$.
Para (1c) tenemos $x_1 = 0.0000000, x_2 = 10.000000, x_3 = 0.14285714$.
Para (1d) tenemos $x_1 = 1.00000000, x_2 = 0.50000000, x_3 = -1.00000000$.
Para (1e) tenemos $x_1 = 0.17682530, x_2 = 0.012692691, x_3 = -0.020654050, x_4 = -1.1826087$.
Para (1f) tenemos $x_1 = 0.78838790, x_2 = -3.1253894, x_3 = 0.1675946, x_4 = 4.5569519$.

15. El algoritmo de pivoteo total con aritmética de precisión simple produce los siguientes resultados.
a. Para (1a) tenemos $x_1 = 9.98, x_2 = 1.00$.
Para (1b) tenemos $x_1 = 1.00, x_2 = 9.98$.
Para (1c) tenemos $x_1 = 0.0724, x_2 = 10.0, x_3 = 0.0952$.
Para (1d) tenemos $x_1 = 0.982, x_2 = 0.500, x_3 = -0.994$.
Para (1e) tenemos $x_1 = 0.161, x_2 = 0.0125, x_3 = -0.0232, x_4 = -1.42$.
Para (1f) tenemos $x_1 = 0.719, x_2 = -2.86, x_3 = 0.146, x_4 = 4.00$.
b. Para (2a) tenemos $x_1 = 10.0, x_2 = 1.00$.
Para (2b) tenemos $x_1 = 1.00, x_2 = 10.0$.
Para (2c) tenemos $x_1 = 0.00, x_2 = 10.0, x_3 = 0.143$.
Para (2d) tenemos $x_1 = 1.01, x_2 = 0.501, x_3 = -1.00$.
Para (2e) tenemos $x_1 = 0.179, x_2 = 0.0127, x_3 = -0.0203, x_4 = -1.15$.
Para (2f) tenemos $x_1 = 0.874, x_2 = -3.49, x_3 = 0.192, x_4 = 5.33$.
c. Para (7a) tenemos $x_1 = 10.000000, x_2 = 1.0000000$.
Para (7b) tenemos $x_1 = 1.0000000, x_2 = 10.000000$.
Para (7c) tenemos $x_1 = 0.0000000, x_2 = 10.000000, x_3 = 0.14285714$.
Para (7d) tenemos $x_1 = 1.00000000, x_2 = 0.50000000, x_3 = -1.00000000$.
Para (7e) tenemos $x_1 = 0.17682530, x_2 = 0.012692691, x_3 = -0.020654050, x_4 = -1.1826087$.
Para (7f) tenemos $x_1 = 0.78838790, x_2 = -3.1253894, x_3 = 0.16759460, x_4 = 4.5569519$.

Conjunto de ejercicios 6.3

1. a. La matriz es singular. **b.** $\begin{bmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{5}{8} & -\frac{1}{8} & -\frac{1}{8} \\ \frac{1}{8} & -\frac{5}{8} & \frac{3}{8} \end{bmatrix}$ **c.** La matriz es singular. **d.** La matriz es singular.

e. $\begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ -\frac{3}{14} & \frac{1}{7} & 0 & 0 \\ \frac{3}{28} & -\frac{11}{7} & 1 & 0 \\ -\frac{1}{2} & 1 & -1 & 1 \end{bmatrix}$ **f.** $\begin{bmatrix} 1 & 0 & 1 & -1 \\ -1 & \frac{5}{3} & \frac{5}{3} & -1 \\ -1 & \frac{2}{3} & \frac{2}{3} & 0 \\ 0 & -\frac{1}{3} & -\frac{4}{3} & 1 \end{bmatrix}$.

3. Las soluciones del sistema lineal obtenidas en los incisos (a) y (b) son, de derecha a izquierda,

$$3,\ -6,\ -2,\ -1 \quad \text{y} \quad 1, 1, 1, 1.$$

5. a. Suponga que $\tilde{A}$ y $\hat{A}$ son inversas de A. Entonces $A\tilde{A} = \tilde{A}A = I$ y $A\hat{A} = \hat{A}A = I$. Por tanto,

$$\tilde{A} = \tilde{A}I = \tilde{A}(A\hat{A}) = (\tilde{A}A)\,\hat{A} = I\hat{A} = \hat{A}.$$

b. $(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$ y $(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)\,B = B^{-1}IB = B^{-1}B = I$, así que $(AB)^{-1} = B^{-1}A^{-1}$ porque hay solamente una inversa.

c. Dado que $A^{-1}A = AA^{-1} = I$, se deduce que A^{-1} es no singular. Puesto que la inversa es singular, tenemos que $(A^{-1})^{-1} = A$.

7. a. Si $C = AB$, donde A y B son triangulares, entonces $a_{ik} = 0$ si $k > i$ y $b_{kj} = 0$ si $k < j$. Por tanto,

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=j}^{i} a_{ik}b_{kj},$$

que tendrá la suma cero salvo que $j \le i$. Por tanto, C es triangular inferior.

b. Tenemos $a_{ik} = 0$ si $k < i$ y $b_{kj} = 0$ si $k > j$. Los pasos se parecen a los del inciso (a).

c. Sea L una matriz triangular inferior no singular. Si queremos resolver la i-ésima columna de L^{-1}, resuelva n sistemas lineales de la forma

$$\begin{bmatrix} l_{11} & 0 & \cdots & \cdots & \cdots & 0 \\ l_{21} & l_{22} & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ l_{i1} & l_{i2} & \cdots & l_{ii} & \ddots & \vdots \\ \vdots & \vdots & & & \ddots & 0 \\ l_{n1} & l_{n2} & \cdots & \cdots & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

donde 1 aparece en la i-ésima posición, para obtener la i-ésima columna de L^{-1}.

9. Las respuestas son las mismas que las del ejercicio 1.

11. a.

$$A^2 = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \\ \frac{1}{6} & 0 & 0 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A^4 = A, \quad A^5 = A^2, \quad A^6 = I, \ldots$$

b.

	Año 1	Año 2	Año 3	Año 4
Edad 1	6000	36000	12000	6000
Edad 2	6000	3000	18000	6000
Edad 3	6000	2000	1000	6000

c.

$$A^{-1} = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \\ \frac{1}{6} & 0 & 0 \end{bmatrix}.$$

La i, j-ésima entrada es el número de escarabajos de edad i que se necesitan para engendrar un escarabajo de edad j.

13. a. Tenemos

$$\begin{bmatrix} 7 & 4 & 4 & 0 \\ -6 & -3 & -6 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2(x_0 - x_1) + \alpha_0 + \alpha_1 \\ 3(x_1 - x_0) - \alpha_1 - 2\alpha_0 \\ \alpha_0 \\ x_0 \end{bmatrix} = \begin{bmatrix} 2(x_0 - x_1) + 3\alpha_0 + 3\alpha_1 \\ 3(x_1 - x_0) - 3\alpha_1 - 6\alpha_0 \\ 3\alpha_0 \\ x_0] \end{bmatrix}$$

b. $B = A^{-1} = \begin{bmatrix} -1 & -\frac{4}{3} & -\frac{4}{3} & 0 \\ 2 & \frac{7}{3} & 2 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Conjunto de ejercicios 6.4

1. Los determinantes de las matrices son **a.** -8 **b.** 14 **c.** 0 **d.** 3.
3. Tenemos $\det A = -5.5$, $\det B = -6$, y $\det AB = \det BA = 33$.
5. $\alpha = -\frac{3}{2}$ y $\alpha = 2$
7. $\alpha = -5$
11. a. La solución es $x_1 = 0$, $x_2 = 10$ y $x_3 = 26$.
    b. Tenemos $D_1 = -1$, $D_2 = 3$, $D_3 = 7$ y $D = 0$ y no hay soluciones.
    c. Tenemos $D_1 = D_2 = D_3 = D = 0$ y hay una cantidad infinita de soluciones.
    e. La regla de Cramer requiere 39 multiplicaciones/divisiones y 20 adiciones/sustracciones.

Conjunto de ejercicios 6.5

1. a. $x_1 = -3$, $x_2 = 3$, $x_3 = 1$
   b. $x_1 = \frac{1}{2}$, $x_2 = -\frac{9}{2}$, $x_3 = \frac{7}{2}$
3. a. $L = \begin{bmatrix} 1 & 0 & 0 \\ 1.5 & 1 & 0 \\ 1.5 & 1 & 1 \end{bmatrix}$ y $U = \begin{bmatrix} 2 & -1 & 1 \\ 0 & 4.5 & 7.5 \\ 0 & 0 & -4 \end{bmatrix}$
   b. $L = \begin{bmatrix} 1 & 0 & 0 \\ -2.106719 & 1 & 0 \\ 3.067193 & 1.197756 & 1 \end{bmatrix}$ y $U = \begin{bmatrix} 1.012 & -2.132 & 3.104 \\ 0 & -0.3955257 & -0.4737443 \\ 0 & 0 & -8.939141 \end{bmatrix}$

c. $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 1 & -1.33333 & 2 & 1 \end{bmatrix}$ y $U = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

d. $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1.849190 & 1 & 0 & 0 \\ -0.4596433 & -0.2501219 & 1 & 0 \\ 2.768661 & -0.3079435 & -5.352283 & 1 \end{bmatrix}$

y

$U = \begin{bmatrix} 2.175600 & 4.023099 & -2.173199 & 5.196700 \\ 0 & 13.43947 & -4.018660 & 10.80698 \\ 0 & 0 & -0.8929510 & 5.091692 \\ 0 & 0 & 0 & 12.03614 \end{bmatrix}$

5 a. $P^tLU = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 \\ 0 & 2 & 3 \\ 0 & 0 & \frac{5}{2} \end{bmatrix}$

b. $P^tLU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -5 & 6 \\ 0 & 0 & 4 \end{bmatrix}$

c. $P^tLU = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 3 & 0 \\ 0 & 5 & -2 & 1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 3 \end{bmatrix}$

d. $P^tLU = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 3 & 0 \\ 0 & 5 & -3 & -1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

7. c.

	Multiplicaciones/divisiones	Adiciones/sustracciones
Factorizando en *LU*	$\frac{1}{3}n^3 - \frac{1}{3}n$	$\frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n$
Resolviendo $Ly = b$	$\frac{1}{2}n^2 - \frac{1}{2}n$	$\frac{1}{2}n^2 - \frac{1}{2}n$
Resolviendo $Ux = y$	$\frac{1}{2}n^2 + \frac{1}{2}n$	$\frac{1}{2}n^2 - \frac{1}{2}n$
Total	$\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$	$\frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n$

d.

	Multiplicaciones/divisiones	Adiciones/sustracciones
Factorizando en *LU*	$\frac{1}{3}n^3 - \frac{1}{3}n$	$\frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n$
Resolviendo $Ly^{(k)} = b^{(k)}$	$(\frac{1}{2}n^2 - \frac{1}{2}n)m$	$(\frac{1}{2}n^2 - \frac{1}{2}n)m$
Resolviendo $Ux^{(k)} = y^{(k)}$	$(\frac{1}{2}n^2 + \frac{1}{2}n)m$	$(\frac{1}{2}n^2 - \frac{1}{2}n)m$
Total	$\frac{1}{3}n^3 + mn^2 - \frac{1}{3}n$	$\frac{1}{3}n^3 + (m - \frac{1}{2})n^2 - (m - \frac{1}{6})n$

Conjunto de ejercicios 6.6

1. (i) Las matrices simétricas están en (a), (b) y (f).
(ii) Las matrices singulares están en (c) y (h).
(iii) Las matrices estrictamente dominantes en forma diagonal están en (a), (b), (c) y (d).
(iv) Las matrices positivas definidas están en (a) y (f).

3. El algoritmo de Choleski produce los siguientes resultados.

a. $L = \begin{bmatrix} 1.414213 & 0 & 0 \\ -0.7071069 & 1.224743 & 0 \\ 0 & -0.8164972 & 1.154699 \end{bmatrix}$

b. $L = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0.5 & 1.658311 & 0 & 0 \\ 0.5 & -0.7537785 & 1.087113 & 0 \\ 0.5 & 0.4522671 & 0.08362442 & 1.240346 \end{bmatrix}$

c. $L = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0.5 & 1.658311 & 0 & 0 \\ -0.5 & -0.4522671 & 2.132006 & 0 \\ 0 & 0 & 0.9380833 & 1.766351 \end{bmatrix}$

d. $L = \begin{bmatrix} 2.449489 & 0 & 0 & 0 \\ 0.8164966 & 1.825741 & 0 & 0 \\ 0.4082483 & 0.3651483 & 1.923538 & 0 \\ -0.4082483 & 0.1825741 & -0.4678876 & 1.606574 \end{bmatrix}$

5. El algoritmo modificado de Choleski produce los siguientes resultados.

a. $x_1 = 1, x_2 = -1, x_3 = 0$

b. $x_1 = 0.2, x_2 = -0.2, x_3 = -0.2, x_4 = 0.25$

c. $x_1 = 1, x_2 = 2, x_3 = -1, x_4 = 2$

d. $x_1 = 0.85863874, x_2 = 2.4188482, x_3 = -0.95811518, x_4 = -1.2722513$

7. Tenemos $x_i = 1$ para toda $i = 1, \ldots, 10$.

9. Sólo la matriz en (d) es positiva definida.

11. $-2 < \alpha < \frac{3}{2}$

13. $0 < \beta < 1$ y $3 < \alpha < 5 - \beta$

15. a. No, considere, por ejemplo, $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

b. Sí, porque $A = A^t$.

c. Sí, porque $\mathbf{x}^t(A + B)\mathbf{x} = \mathbf{x}^tA\mathbf{x} + \mathbf{x}^tB\mathbf{x}$.

d. Sí, porque $\mathbf{x}^tA^2\mathbf{x} = \mathbf{x}^tA^tA\mathbf{x} = (A\mathbf{x})^t(A\mathbf{x}) \geq 0$ y como A es no singular, la igualdad se mantiene sólo si $\mathbf{x} = \mathbf{0}$.

e. No, considere, por ejemplo, $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ y $B = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$.

17. a. Dado que $\det A = 3\alpha - 2\beta$, A es singular si y sólo si $\alpha = 2\beta/3$. b. $|\alpha| > 1$, $|\beta| < 1$ c. $\beta = 1$

d. $\alpha > \frac{2}{3}$, $\beta = 1$

19. Un ejemplo es $A = \begin{bmatrix} 1.0 & 0.2 \\ 0.1 & 1.0 \end{bmatrix}$.

23. $i_1 = 0.6785047$, $i_2 = 0.4214953$, $i_3 = 0.2570093$, $i_4 = 0.1542056$, $i_5 = 0.1028037$

25. a. Al aparearse el macho i con la hembra j, se engendra una cría con las mismas características de ala que cuando se aparea un macho j con una hembra i.

b. No. Considere, por ejemplo, $\mathbf{x} = (1, 0, -1)^t$.

Conjunto de ejercicios 7.1

1. a. Tenemos $\|\mathbf{x}\|_\infty = 4$ y $\|\mathbf{x}\|_2 = 5.220153$.

b. Tenemos $\|\mathbf{x}\|_\infty = 4$ y $\|\mathbf{x}\|_2 = 5.477226$.

c. Tenemos $\|\mathbf{x}\|_\infty = 2^k$ y $\|\mathbf{x}\|_2 = (1 + 4^k)^{1/2}$

d. Tenemos $\|\mathbf{x}\|_\infty = 4/(k + 1)$ y $\|\mathbf{x}\|_2 = (16/(k + 1)^2 + 4/k^4 + k^4e^{-2k})^{1/2}$.

3. a. Tenemos $\lim_{k\to\infty} \mathbf{x}^{(k)} = (0, 0, 0)^t$.

b. Tenemos $\lim_{k\to\infty} \mathbf{x}^{(k)} = (0, 1, 3)^t$.

c. Tenemos $\lim_{k\to\infty} \mathbf{x}^{(k)} = (0, 0, \frac{1}{2})^t$.

d. Tenemos $\lim_{k\to\infty} \mathbf{x}^{(k)} = (1, -1, 1)^t$.

5. a. Tenemos $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty = 8.57 \times 10^{-4}$ y $\|A\hat{\mathbf{x}} - \mathbf{b}\|_\infty = 2.06 \times 10^{-4}$.

b. Tenemos $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty = 0.90$ y $\|A\hat{\mathbf{x}} - \mathbf{b}\|_\infty = 0.27$.

c. Tenemos $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty = 0.5$ y $\|A\hat{\mathbf{x}} - \mathbf{b}\|_\infty = 0.3$.

d. Tenemos $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty = 6.55 \times 10^{-2}$ y $\|A\hat{\mathbf{x}} - \mathbf{b}\|_\infty = 0.32$.

7. Sea $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ y $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. Entonces $\|AB\|_\otimes = 2$, pero $\|A\|_\otimes \cdot \|B\|_\otimes = 1$.

9. b. Tenemos

(4a) $\|A\|_F = \sqrt{326}$

(4b) $\|A\|_F = \sqrt{326}$

(4c) $\|A\|_F = 4$

(4d) $\|A\|_F = 4\sqrt{148}$.

Conjunto de ejercicios 7.2

1. a. El valor característico $\lambda_1 = 3$ tiene el vector característico $\mathbf{x}_1 = (1, -1)^t$, y el valor característico $\lambda_2 = 1$ tiene el vector característico $\mathbf{x}_2 = (1, 1)^t$.

b. El valor característico $\lambda_1 = \frac{1+\sqrt{5}}{2}$ tiene el vector característico $\mathbf{x}_1 = \left(1, \frac{1+\sqrt{5}}{2}\right)^t$, y el valor característico $\lambda_2 = \frac{1-\sqrt{5}}{2}$ tiene el vector característico $\mathbf{x}_2 = \left(1, \frac{1-\sqrt{5}}{2}\right)^t$.

c. El valor característico $\lambda_1 = \frac{1}{2}$ tiene el vector característico $\mathbf{x}_1 = (1, 1)^t$, y el valor característico $\lambda_2 = -\frac{1}{2}$ tiene el vector característico $\mathbf{x}_2 = (1, -1)^t$.

d. El valor característico $\lambda_1 = 0$ tiene el vector característico $\mathbf{x}_1 = (1, -1)^t$, y el valor característico $\lambda_2 = -1$ tiene el vector característico $\mathbf{x}_2 = (1, -2)^t$.

e. El valor característico $\lambda_1 = \lambda_2 = 3$ tiene los vectores característicos $\mathbf{x}_1 = (0, 0, 1)^t$ y $\mathbf{x}_2 = (1, 1, 0)^t$, y el valor característico $\lambda_3 = 1$ tiene el vector característico $\mathbf{x}_3 = (1, -1, 0)^t$.

f. El valor característico $\lambda_1 = 7$ tiene el vector característico $\mathbf{x}_1 = (1, 4, 4)^t$, el valor característico $\lambda_2 = 3$ tiene el vector característico $\mathbf{x}_2 = (1, 2, 0)^t$, y el valor característico $\lambda_3 = -1$ tiene el vector característico $\mathbf{x}_3 = (1, 0, 0)^t$.

g. El valor característico $\lambda_1 = \lambda_2 = 1$ tiene los vectores característicos $\mathbf{x}_1 = (-1, 1, 0)^t$, $\mathbf{x}_2 = (-1, 0, 1)^t$, y el valor característico $\lambda_3 = 5$ tiene el vector característico $\mathbf{x}_3 = (1, 2, 1)^t$.

h. El valor característico $\lambda_1 = 3$ tiene el vector característico $\mathbf{x}_1 = (-1, 1, 2)^t$, el valor característico $\lambda_2 = 4$ tiene el vector característico $\mathbf{x}_2 = (0, 1, 2)^t$, y el valor característico $\lambda_3 = -2$ tiene el vector característico $\mathbf{x}_3 = (-3, 8, 1)^t$.

3. Sólo la matriz en (c) es convergente.

5. a. 3 **b.** 1.618034 **c.** 0.5 **d.** 3.162278 **e.** 3 **f.** 8.224257 **g.** 5.203527 **h.** 5.601152

9. a. $\det(A - \lambda I) = \det((A - \lambda I)^t) = \det(A^t - \lambda I)$
   b. Si $A\mathbf{x} = \lambda\mathbf{x}$, entonces $A^2\mathbf{x} = \lambda A\mathbf{x} = \lambda^2\mathbf{x}$, y por inducción, $A^k\mathbf{x} = \lambda^k\mathbf{x}$.
   c. Si $A\mathbf{x} = \lambda\mathbf{x}$ y A^{-1} existen, entonces $\mathbf{x} = \lambda A^{-1}\mathbf{x}$. Por el ejercicio 8(b), $\lambda \neq 0$, así, $\frac{1}{\lambda}\mathbf{x} = A^{-1}\mathbf{x}$.
   d. Puesto que $A^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}$, tenemos $(A^{-1})^2\mathbf{x} = \frac{1}{\lambda}A^{-1}\mathbf{x} = \frac{1}{\lambda^2}\mathbf{x}$. La inducción matemática da

$$(A^{-1})^k\mathbf{x} = \frac{1}{\lambda^k}\mathbf{x}.$$

   e. Si $A\mathbf{x} = \lambda\mathbf{x}$, entonces

$$q(A)\mathbf{x} = q_0\mathbf{x} + q_1A\mathbf{x} + \ldots + q_kA^k\mathbf{x} = q_0\mathbf{x} + q_1\lambda\mathbf{x} + \ldots + q_k\lambda^k\mathbf{x} = q(\lambda)\mathbf{x}.$$

   f. Sea $A - \alpha I$ no singular. Puesto que $A\mathbf{x} = \lambda\mathbf{x}$,

$$(A - \alpha I)\mathbf{x} = A\mathbf{x} - \alpha I\mathbf{x} = \lambda\mathbf{x} - \alpha\mathbf{x} = (\lambda - \alpha)\mathbf{x}.$$

Por tanto,

$$\frac{1}{\lambda - \alpha}\mathbf{x} = (A - \alpha I)^{-1}\mathbf{x}.$$

11. a. Tenemos el valor característico real $\lambda = 1$ con el vector característico $\mathbf{x} = (6, 3, 1)^t$.
    b. Seleccione cualquier múltiplo del vector $(6, 3, 1)^t$.
13. Sea $A\mathbf{x} = \lambda\mathbf{x}$. Entonces $|\lambda|\,\|\mathbf{x}\| = \|A\mathbf{x}\| \leq \|A\|\,\|\mathbf{x}\|$, que implica $|\lambda| \leq \|A\|$. También, $(1/\lambda)\mathbf{x} = A^{-1}\mathbf{x}$, por lo que $1/|\lambda| \leq \|A^{-1}\|$ y $\|A^{-1}\|^{-1} \leq |\lambda|$.

Conjunto de ejercicios 7.3

1. Dos iteraciones del método de Jacobi producen los siguientes resultados.
   a. $(0.1428571, -0.3571429, 0.4285714)^t$
   b. $(0.97, 0.91, 0.74)^t$
   c. $(-0.65, 1.65, -0.4, -2.475)^t$
   d. $(-0.5208333, -0.04166667, -0.2166667, 0.4166667)^t$
   e. $(1.325, -1.6, 1.6, 1.675, 2.425)^t$
   f. $(0.6875, 1.125, 0.6875, 1.375, 0.5625, 1.375)^t$
3. El algoritmo de Jacobi produce los siguientes resultados.
   a. $x^{(10)} = (0.03507839, -0.2369262, 0.6578015)^t$
   b. $x^{(6)} = (0.9957250, 0.9577750, 0.7914500)^t$
   c. $x^{(22)} = (-0.7975853, 2.794795, -0.2588888, -2.251879)^t$
   d. $x^{(14)} = (-0.7529267, 0.04078538, -0.2806091, 0.6911662)^t$
   e. $x^{(12)} = (0.7870883, -1.003036, 1.866048, 1.912449, 1.985707)^t$
   f. $x^{(17)} = (0.9996805, 1.999774, 0.9996805, 1.999840, 0.9995482, 1.999840)^t$
5. Dos iteraciones del método SOR producen los siguientes resultados.
   a. $(0.05410079, -0.2115435, 0.6477159)^t$
   b. $(0.9876790, 0.9784935, 0.7899328)^t$
   c. $(-0.71885, 2.818822, -0.2809726, -2.235422)^t$
   d. $(-0.6604902, 0.03700749, -0.2493513, 0.6561139)^t$
   e. $(1.079675, -1.260654, 2.042489, 1.995373, 2.049536)^t$
   f. $(0.8318750, 1.647766, 0.9189856, 1.791281, 0.8712129, 1.959155)^t$

7. El algoritmo SOR produce los siguientes resultados.

a. $\mathbf{x}^{(12)} = (0.03488469, -0.2366474, 0.6579013)^t$

b. $\mathbf{x}^{(7)} = (0.9958341, 0.9579041, 0.7915756)^t$

c. $\mathbf{x}^{(8)} = (-0.7976009, 2.795288, -0.2588293, -2.251768)^t$

d. $\mathbf{x}^{(7)} = (-0.7534489, 0.04106617, -0.2808146, 0.6918049)^t$

e. $\mathbf{x}^{(10)} = (0.7866310, -1.002807, 1.866530, 1.912645, 1.989792)^t$

f. $\mathbf{x}^{(7)} = (0.9999442, 1.999934, 1.000033, 1.999958, 0.9999815, 2.000007)^t$

9. a.

$$T_j = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & -1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \text{ y } \det(\lambda I - T_j) = \lambda^3 + \frac{5}{4}x.$$

Así los valores característicos de T_j son 0 y $\pm\frac{\sqrt{5}}{2}i$, por lo que $\rho(T_j) = \frac{\sqrt{5}}{2} > 1$.

b. $\mathbf{x}^{(25)} = (-20.827873, 2.0000000, -22.827873)^t$

c.

$$T_g = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} \end{bmatrix} \text{ y } \det(\lambda I - T_g) = \lambda\left(\lambda + \frac{1}{2}\right)^2.$$

Así, los valores característicos de T_g son 0, $-1/2$ y $-1/2$; y $\rho(T_g) = 1/2$.

d. $\mathbf{x}^{(23)} = (1.0000023, 1.9999975, -1.0000001)^t$ no sobrepasa 10^{-5} en la norma l_∞.

11. a. Reste $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ a $\mathbf{x}^{(k)} = T\mathbf{x}^{(k-1)} + \mathbf{c}$ para obtener $\mathbf{x}^{(k)} - \mathbf{x} = T(\mathbf{x}^{(k-1)} - \mathbf{x})$. Por tanto,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \|T\| \ \|\mathbf{x}^{(k-1)} - \mathbf{x}\|.$$

Por inducción, obtenemos

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \|T\|^k \ \|\mathbf{x}^{(0)} - \mathbf{x}\|.$$

El resto de la demostración es similar a la del corolario 2.5.

b. La última columna no tiene elementos cuando $\|T\|_\infty = 1$.

	$\|\mathbf{x}^{(2)} - \mathbf{x}\|_\infty$	$\|T\|_\infty$	$\|T\|_\infty^2 \|\mathbf{x}^{(0)} - \mathbf{x}\|_\infty$	$\frac{\|T\|_\infty^2}{1-\|T\|_\infty}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty$
1(a)	0.22932	0.857143	0.48335	2.9388
1(b)	0.051579	0.3	0.089621	0.11571
1(c)	1.1453	0.9	2.2642	20.25
1(d)	0.27511	1	0.75342	
1(e)	0.59743	1	1.9897	
1(f)	0.875	0.75	1.125	3.375

15.

	33 iteraciones de Jacobi	8 iteraciones de Gauss-Seidel	13 iteraciones SOR ($\omega = 1.2$)
x_1	1.53873501	1.53873270	1.53873549
x_2	0.73142167	0.73141966	0.73142226
x_3	0.10797136	0.10796931	0.10797063
x_4	0.17328530	0.17328340	0.17328480
x_5	0.04055865	0.04055595	0.04055737
x_6	0.08525019	0.08524787	0.08524925
x_7	0.16645040	0.16644711	0.16644868
x_8	0.12198156	0.12197878	0.12198026
x_9	0.10125265	0.10124911	0.10125043
x_{10}	0.09045966	0.09045662	0.09045793
x_{11}	0.07203172	0.07202785	0.07202912
x_{12}	0.07026597	0.07026266	0.07026392
x_{13}	0.06875835	0.06875421	0.06875546
x_{14}	0.06324659	0.06324307	0.06324429
x_{15}	0.05971510	0.05971083	0.05971200
x_{16}	0.05571199	0.05570834	0.05570949
x_{17}	0.05187851	0.05187416	0.05187529
x_{18}	0.04924911	0.04924537	0.04924648
x_{19}	0.04678213	0.04677776	0.04677885
x_{20}	0.04448679	0.04448303	0.04448409
x_{21}	0.04246924	0.04246493	0.04246597
x_{22}	0.04053818	0.04053444	0.04053546
x_{23}	0.03877273	0.03876852	0.03876952
x_{24}	0.03718190	0.03717822	0.03717920
x_{25}	0.03570858	0.03570451	0.03570548
x_{26}	0.03435107	0.03434748	0.03434844
x_{27}	0.03309542	0.03309152	0.03309246
x_{28}	0.03192212	0.03191866	0.03191958
x_{29}	0.03083007	0.03082637	0.03082727
x_{30}	0.02980997	0.02980666	0.02980755
x_{31}	0.02885510	0.02885190	0.02885248
x_{32}	0.02795937	0.02795621	0.02795707
x_{33}	0.02711787	0.02711458	0.02711543
x_{34}	0.02632478	0.02632179	0.02632262

	33 iteraciones de Jacobi	8 iteraciones de Gauss-Seidel	13 iteraciones SOR ($\omega = 1.2$)
x_{35}	0.02557705	0.02557397	0.02557479
x_{36}	0.02487017	0.02486733	0.02486814
x_{37}	0.02420147	0.02419858	0.02419938
x_{38}	0.02356750	0.02356482	0.02356560
x_{39}	0.02296603	0.02296333	0.02296410
x_{40}	0.02239424	0.02239171	0.02239247
x_{41}	0.02185033	0.02184781	0.02184855
x_{42}	0.02133203	0.02132965	0.02133038
x_{43}	0.02083782	0.02083545	0.02083615
x_{44}	0.02036585	0.02036360	0.02036429
x_{45}	0.01991483	0.01991261	0.01991324
x_{46}	0.01948325	0.01948113	0.01948175
x_{47}	0.01907002	0.01906793	0.01906846
x_{48}	0.01867387	0.01867187	0.01867239
x_{49}	0.01829386	0.01829190	0.01829233
x_{50}	0.71792896	0.01792707	0.01792749
x_{51}	0.01757833	0.01757648	0.01757683
x_{52}	0.01724113	0.01723933	0.01723968
x_{53}	0.01691660	0.01691487	0.01691517
x_{54}	0.01660406	0.01660237	0.01660267
x_{55}	0.01630279	0.01630127	0.01630146
x_{56}	0.01601230	0.01601082	0.01601101
x_{57}	0.01573198	0.01573087	0.01573077
x_{58}	0.01546129	0.01546020	0.01546010
x_{59}	0.01519990	0.01519909	0.01519878
x_{60}	0.01494704	0.01494626	0.01494595
x_{61}	0.01470181	0.01470085	0.01470077
x_{62}	0.01446510	0.01446417	0.01446409
x_{63}	0.01423556	0.01423437	0.01423461
x_{64}	0.01401350	0.01401233	0.01401256
x_{65}	0.01380328	0.01380234	0.01380242
x_{66}	0.01359448	0.01359356	0.01359363
x_{67}	0.01338495	0.01338434	0.01338418
x_{68}	0.01318840	0.01318780	0.01318765
x_{69}	0.01297174	0.01297109	0.01297107
x_{70}	0.01278663	0.01278598	0.01278597
x_{71}	0.01270328	0.01270263	0.01270271
x_{72}	0.01252719	0.01252656	0.01252663
x_{73}	0.01237700	0.01237656	0.01237654
x_{74}	0.01221009	0.01220965	0.01220963
x_{75}	0.01129043	0.01129009	0.01129008
x_{76}	0.01114138	0.01114104	0.01114102
x_{77}	0.01217337	0.01217312	0.01217313
x_{78}	0.01201771	0.01201746	0.01201746
x_{79}	0.01542910	0.01542896	0.01542896
x_{80}	0.01523810	0.01523796	0.01523796

Conjunto de ejercicios 7.4

1. El número de condición en $\|\cdot\|_\infty$ es: **a.** 50 **b.** 241.37 **c.** 600,002 **d.** 339,866 **e.** 12 **h.** 198.17

3. La matriz está mal condicionada porque $K_\infty = 60002$. Tenemos $\tilde{\mathbf{x}} = (-1.0000, 2.0000)^t$.

5. a. Tenemos $\tilde{\mathbf{x}} = (188.9998, 92.99998, 45.00001, 27.00001, 21.00002)^t$.

b. El número de condición es $K_\infty = 80$.

c. La solución exacta es $\mathbf{x} = (189, 93, 45, 27, 21)^t$.

9. Para la matriz de 3×3 de Hilbert tenemos

$$\hat{H}^{-1} = \begin{bmatrix} 8.968 & -35.77 & 29.77 \\ -35.77 & 190.6 & -178.6 \\ 29.77 & -178.6 & 178.6 \end{bmatrix}, \qquad \hat{H} = \begin{bmatrix} 0.9799 & 0.4870 & 0.3238 \\ 0.4860 & 0.3246 & 0.2434 \\ 0.3232 & 0.2433 & 0.1949 \end{bmatrix},$$

y $\|H - \hat{H}\|_\infty = 0.04260$.

Conjunto de ejercicios 7.5

1. $(0.18, 0.13)^t$

b. $(0.19, 0.10)^t$

c. La eliminación gaussiana da la mejor respuesta, pues $\mathbf{v}^{(2)} = (0, 0)^t$ en el método de gradiente conjugado.

d. $(0.13, 0.21)^t$. No hay mejora, aunque $\mathbf{v}^{(2)} \neq \mathbf{0}$

3. a. $(1.00, -1.00, 1.00)^t$

b. $(0.827, 0.0453, -0.0357)^t$

c. El pivoteo parcial y el parcial escalado también dan $(1.00, -1.00, 1.00)^t$.

d. $(0.776, 0.238, -0.185)^t$;

El residuo de (3b) es $(-0.0004, -0.0038, 0.0037)^t$, y el residuo del inciso (3d) es $(0.0022, -0.0038, 0.0024)^t$. No parece haber mucha mejora, si acaso hay alguna. El error de redondeo está más presente por la mayor cantidad de multiplicaciones de matriz.

5. a. $\mathbf{x}^{(2)} = (0.1535933456, -0.1697932117, 0.5901172091)^t$, $\|\mathbf{r}^{(2)}\|_\infty = 0.221$.

b. $\mathbf{x}^{(2)} = (0.9993129510, 0.9642734456, 0.7784266575)^t$, $\|\mathbf{r}^{(2)}\|_\infty = 0.144$.

c. $\mathbf{x}^{(2)} = (-0.7290954114, 2.515782452, -0.6788904058, -2.331943982)^t$, $\|\mathbf{r}^{(2)}\|_\infty = 2.2$.

d. $\mathbf{x}^{(2)} = (-0.7071108901, -0.0954748881, -0.3441074093, 0.5256091497)^t$, $\|\mathbf{r}^{(2)}\|_\infty = 0.39$.

e. $\mathbf{x}^{(2)} = (0.5335968381, 0.9367588935, 1.339920949, 1.743083004, 1.743083004)^t$, $\|\mathbf{r}^{(2)}\|_\infty = 1.3$.

f. $\mathbf{x}^{(2)} = (1.022375671, 1.686451893, 1.022375671, 2.060919568, 0.8310997764, 2.060919568)^t$, $\|\mathbf{r}^{(2)}\|_\infty = 1.13$.

7. a. $\mathbf{x}^{(3)} = (0.06185567013, -0.1958762887, 0.6185567010)^t$, $\|\mathbf{r}^{(3)}\|_\infty = 0.4 \times 10^{-9}$.

b. $\mathbf{x}^{(3)} = (0.9957894738, 0.9578947369, 0.7915789474)^t$, $\|\mathbf{r}^{(3)}\|_\infty = 0.1 \times 10^{-9}$.

c. $\mathbf{x}^{(4)} = (-0.7976470579, 2.795294120, -0.2588235305, -2.251764706)^t$, $\|\mathbf{r}^{(4)}\|_\infty = 0.39 \times 10^{-7}$.

d. $\mathbf{x}^{(4)} = (-0.7534246575, 0.04109589039, -0.2808219179, 0.6917808219)^t$, $\|\mathbf{r}^{(4)}\|_\infty = 0.11 \times 10^{-9}$.

e. $\mathbf{x}^{(5)} = (0.4516129032, 0.7096774197, 1.677419355, 1.741935483, 1.806451613)^t$, $\|\mathbf{r}^{(5)}\|_\infty = 0.2 \times 10^{-9}$.

f. $\mathbf{x}^{(4)} = (1.000000000, 2.000000000, 1.000000000, 2.000000000, 0.9999999997, 2.000000000)^t$,

$\|r^{(4)}\|_\infty = 0.44 \times 10^{-9}$.

9.

a.	49 iteraciones de Jacobi	28 iteraciones de Gauss-Seidel	13 iteraciones SOR ($\omega = 1.3$)	9 iteraciones de gradiente conjugado
x_1	0.93406183	0.93406917	0.93407584	0.93407713
x_2	0.97473885	0.97475285	0.97476180	0.97476363
x_3	1.10688692	1.10690302	1.10691093	1.10691243
x_4	1.42346150	1.42347226	1.42347591	1.42347699
x_5	0.85931331	0.85932730	0.85933633	0.85933790
x_6	0.80688119	0.80690725	0.80691961	0.80692197
x_7	0.85367746	0.85370564	0.85371536	0.85372011
x_8	1.10688692	1.10690579	1.10691075	1.10691250
x_9	0.87672774	0.87674384	0.87675177	0.87675250
x_{10}	0.80424512	0.80427330	0.80428301	0.80428524
x_{11}	0.80688119	0.80691173	0.80691989	0.80692252
x_{12}	0.97473885	0.97475850	0.97476265	0.97476392
x_{13}	0.93003466	0.93004542	0.93004899	0.93004987
x_{14}	0.87672774	0.87674661	0.87675155	0.87675298
x_{15}	0.85931331	0.85933296	0.85933709	0.85933979
x_{16}	0.93406183	0.93407462	0.93407672	0.93407768

b.	60 iteraciones de Jacobi	35 iteraciones de Gauss-Seidel	23 iteraciones SOR ($\omega = 1.2$)	11 iteraciones de gradiente conjugado
x_1	0.39668038	0.39668651	0.39668915	0.39669775
x_2	0.07175540	0.07176830	0.07177348	0.07178516
x_3	−0.23080396	−0.23078609	−0.23077981	−0.23076923
x_4	0.24549277	0.24550989	0.24551535	0.24552253
x_5	0.83405412	0.83406516	0.83406823	0.83407148
x_6	0.51497606	0.51498897	0.51499414	0.51500583
x_7	0.12116003	0.12118683	0.12119625	0.12121212
x_8	−0.24044414	−0.24040991	−0.24039898	−0.24038462
x_9	0.37873579	0.37876891	0.37877812	0.37878788
x_{10}	1.09073364	1.09075392	1.09075899	1.09076341
x_{11}	0.54207872	0.54209658	0.54210286	0.54211344
x_{12}	0.13838259	0.13841682	0.13842774	0.13844211
x_{13}	−0.23083868	−0.23079452	−0.23078224	−0.23076923
x_{14}	0.41919067	0.41923122	0.41924136	0.41925019
x_{15}	1.15015953	1.15018477	1.15019025	1.15019425
x_{16}	0.51497606	0.51499318	0.51499864	0.51500583
x_{17}	0.12116003	0.12119315	0.12120236	0.12121212
x_{18}	−0.24044414	−0.24040359	−0.24039345	−0.24038462
x_{19}	0.37873579	0.37877365	0.37878188	0.37878788
x_{20}	1.09073364	1.09075629	1.09076069	1.09076341
x_{21}	0.39668038	0.39669142	0.39669449	0.39669775
x_{22}	0.07175540	0.07177567	0.07178074	0.07178516
x_{23}	−0.23080396	−0.23077872	−0.23077323	−0.23076923
x_{24}	0.24549277	0.24551542	0.24551982	0.24552253
x_{25}	0.83405412	0.83406793	0.83407025	0.83407148

c.	15 iteraciones de Jacobi	9 iteraciones de Gauss-Seidel	8 iteraciones SOR ($\omega = 1.2$)	8 iteraciones de gradiente conjugado
x_1	−3.07611424	−3.07611739	−3.07611796	−3.07611794
x_2	−1.65223176	−1.65223563	−1.65223579	−1.65223582
x_3	−0.53282391	−0.53282528	−0.53282531	−0.53282528
x_4	−0.04471548	−0.04471608	−0.04471609	−0.04471604
x_5	0.17509673	0.17509661	0.17509661	0.17509661
x_6	0.29568226	0.29568223	0.29568223	0.29568218
x_7	0.37309012	0.37309011	0.37309011	0.37309011
x_8	0.42757934	0.42757934	0.42757934	0.42757927
x_9	0.46817927	0.46817927	0.46817927	0.46817927
x_{10}	0.49964748	0.49964748	0.49964748	0.49964748
x_{11}	0.52477026	0.52477026	0.52477026	0.52477027
x_{12}	0.54529835	0.54529835	0.54529835	0.54529836
x_{13}	0.56239007	0.56239007	0.56239007	0.56239009
x_{14}	0.57684345	0.57684345	0.57684345	0.57684347
x_{15}	0.58922662	0.58922662	0.58922662	0.58922664
x_{16}	0.59995522	0.59995522	0.59995522	0.59995523
x_{17}	0.60934045	0.60934045	0.60934045	0.60934045
x_{18}	0.61761997	0.61761997	0.61761997	0.61761998
x_{19}	0.62497846	0.62497846	0.62497846	0.62497847
x_{20}	0.63156161	0.63156161	0.63156161	0.63156161
x_{21}	0.63748588	0.63748588	0.63748588	0.63748588
x_{22}	0.64284553	0.64284553	0.64284553	0.64284553
x_{23}	0.64771764	0.64771764	0.64771764	0.64771764
x_{24}	0.65216585	0.65216585	0.65216585	0.65216585
x_{25}	0.65624320	0.65624320	0.65624320	0.65624320
x_{26}	0.65999423	0.65999423	0.65999423	0.65999422
x_{27}	0.66345660	0.66345660	0.66345660	0.66345660
x_{28}	0.66666242	0.66666242	0.66666242	0.66666242
x_{29}	0.66963919	0.66963919	0.66963919	0.66963919
x_{30}	0.67241061	0.67241061	0.67241061	0.67241060
x_{31}	0.67499722	0.67499722	0.67499722	0.67499721
x_{32}	0.67741692	0.67741692	0.67741691	0.67741691
x_{33}	0.67968535	0.67968535	0.67968535	0.67968535
x_{34}	0.68181628	0.68181628	0.68181628	0.68181628
x_{35}	0.68382184	0.68382184	0.68382184	0.68382184
x_{36}	0.68571278	0.68571278	0.68571278	0.68571278
x_{37}	0.68749864	0.68749864	0.68749864	0.68749864
x_{38}	0.68918652	0.68918652	0.68918652	0.68918652
x_{39}	0.69067718	0.69067718	0.69067718	0.69067717
x_{40}	0.68363346	0.68633346	0.68363346	0.68363349

11. a.

Solución	Residual
2.55613420	0.00668246
4.09171393	−0.00533953
4.60840390	−0.01739814
3.64309950	−0.03171624
5.13950533	0.01308093
7.19697808	−0.02081095
7.68140405	−0.04593118
5.93227784	0.01692180
5.81798997	0.04414047
5.85447806	0.03319707
5.94202521	−0.00099947
4.42152959	−0.00072826
3.32211695	0.02363822
4.49411604	0.00982052
4.80968966	0.00846967
3.81108707	−0.01312902

Esto converge en 6 iteraciones con tolerancia de 5.00×10^{-2} en la norma l_∞ y $\|\mathbf{r}^{(6)}\|_\infty = 0.046$.

b.

Solución	Residual
2.55613420	0.00668246
4.09171393	−0.00533953
4.60840390	−0.01739814
3.64309950	−0.03171624
5.13950533	0.01308093
7.19697808	−0.02081095
7.68140405	−0.04593118
5.93227784	0.01692180
5.81798996	0.04414047
5.85447805	0.03319706
5.94202521	−0.00099947
4.42152959	−0.00072826
3.32211694	0.02363822
4.49411603	0.00982052
4.80968966	0.00846967
3.81108707	−0.01312902

Esto converge en 6 iteraciones con tolerancia de 5.00×10^{-2} en la norma l_∞ y $\|\mathbf{r}^{(6)}\|_\infty = 0.046$.

c. Todas las tolerancias conducen a las mismas especificaciones de convergencia.

13. a. Sea $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ un conjunto de vectores A-ortogonales distintos de cero para la matriz simétrica definida positiva A. En tal caso, $\langle \mathbf{v}^{(i)}, A\mathbf{v}^{(j)} \rangle = 0$, si $i \neq j$. Suponga que

$$c_1\mathbf{v}^{(1)} + c_2\mathbf{v}^{(2)} + \cdots + c_n\mathbf{v}^{(n)} = \mathbf{0},$$

donde no todas las c_i son cero. Suponga que k es el entero más pequeño para el cual $c_k \neq 0$. Así,

$$c_k\mathbf{v}^{(k)} + c_{k+1}\mathbf{v}^{(k+1)} + \cdots + c_n\mathbf{v}^{(n)} = \mathbf{0}.$$

Despejamos $\mathbf{v}^{(k)}$ y obtenemos

$$\mathbf{v}^{(k)} = -\frac{c_{k+1}}{c_k}\mathbf{v}^{(k+1)} - \cdots - \frac{c_n}{c_k}\mathbf{v}^{(n)}.$$

Al multiplicar por A, tenemos

$$A\mathbf{v}^{(k)} = -\frac{c_{k+1}}{c_k}A\mathbf{v}^{(k+1)} - \cdots - \frac{c_n}{c_k}A\mathbf{v}^{(n)},$$

por lo que

$$(\mathbf{v}^{(k)})^t A\mathbf{v}^{(k)} = -\frac{c_{k+1}}{c_k}(\mathbf{v}^{(k)})^t A\mathbf{v}^{(k+1)} - \cdots - \frac{c_n}{c_k}(\mathbf{v}^{(k)})^t A\mathbf{v}^{(n)}$$

$$= -\frac{c_{k+1}}{c_k}\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(k+1)}\rangle - \cdots - \frac{c_n}{c_k}\langle \mathbf{v}^{(k)}, A\mathbf{v}^{(n)}\rangle$$

$$= -\frac{c_{k+1}}{c_k}\cdot 0 - \cdots - \frac{c_n}{c_k}\cdot 0.$$

Puesto que A es definida positiva, $\mathbf{v}^{(k)} = \mathbf{0}$, lo que es una contradicción. Así, todas las c_i deben ser cero, y $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ es linealmente independiente.

b. Sea $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ un conjunto de vectores A-ortogonales distintos de cero para la matriz simétrica definida positiva A, y $\mathbf{z}$ ortogonal a $\mathbf{v}^{(i)}$, para cada $i = 1, \ldots, n$. Del inciso (a), el conjunto $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ es linealmente independiente, por lo que hay una colección de constantes $\beta_1, \ldots, \beta_n$ con

$$\mathbf{z} = \sum_{i=1}^{n} \beta_i \mathbf{v}^{(i)}.$$

Por consiguiente,

$$\langle \mathbf{z}, \mathbf{z}\rangle = \mathbf{z}^t\mathbf{z} = \sum_{i=1}^{n} \beta_i \mathbf{z}^t\mathbf{v}^{(i)} = \sum_{i=1}^{n} \beta_i \cdot 0 = 0,$$

y el teorema 7.30, inciso (v), implica que $\mathbf{z} = \mathbf{0}$.

Conjunto de ejercicios 8.1

1. El polinomio lineal de mínimos cuadrados es $1.70784x + 0.89968$.

3. Los polinomios de mínimos cuadrados con sus errores son, respectivamente, $0.6208950 + 1.219621x$, con $E = 2.719 \times 10^{-5}$; $0.5965807 + 1.253293x - 0.01085343x^2$, con $E = 1.801 \times 10^{-5}$; y $0.6290193 + 1.185010x + 0.03533252x^2 - 0.01004723x^3$, con $E = 1.741 \times 10^{-5}$.

5. a. El polinomio lineal de mínimos cuadrados es $72.0845x - 194.138$, con un error de 329.

b. El polinomio de segundo grado de mínimos cuadrados es $6.61821x^2 - 1.14352x + 1.23556$, con un error de 1.44×10^{-3}.

c. El polinomio de tercer grado de mínimos cuadrados es $-0.0136742x^3 + 6.84557x^2 - 2.37919x + 3.42904$, con un error de 5.27×10^{-4}.

d. La aproximación de mínimos cuadrados de la forma be^{ax} es $24.2588e^{0.372382x}$, con un error de 418.

e. La aproximación de mínimos cuadrados de la forma bx^a es $6.23903x^{2.01954}$, con un error de 0.00703.

7. a. $k = 0.8996$, $E(k) = 0.295$ **b.** $k = 0.9052$, $E(k) = 0.128$. El inciso (b) encaja óptimamente en los datos experimentales totales.

9. La línea de mínimos cuadrados para el punto promedio es 0.101 (puntuación ACT) + 0.487.

11. El polinomio lineal de mínimos cuadrados produce $y \approx 0.17952x + 8.2084$.

13. a. $\ln R = \ln 1.304 + 0.5756 \ln W$

b. $E = 25.25$

c. $\ln R = \ln 1.051 + 0.7006 \ln W + 0.06695(\ln W)^2$

d. $E = \sum_{i=1}^{37} \left(R_i - bW_i^a e^{c(\ln W_i)^2}\right)^2 = 20.30$

Conjunto de ejercicios 8.2

1. Las aproximaciones lineales de mínimos cuadrados son:

a. $P_1(x) = 1.833333 + 4x$

b. $P_1(x) = -1.600003 + 3.600003x$

c. $P_1(x) = 1.140981 - 0.2958375x$

d. $P_1(x) = 1.1945267 + 3.000001x$

e. $P_1(x) = 0.6109245 + 0.09167105x$

f. $P_1(x) = -1.861455 + 1.666667x$

3. Las aproximaciones lineales de mínimos cuadrados en $[-1, 1]$ son

a. $P_1(x) = 3.333333 - 2x$

b. $P_1(x) = 0.6000025x$

c. $P_1(x) = 0.5493063 - 0.2958375x$

d. $P_1(x) = 1.175201 + 1.103639x$

e. $P_1(x) = 0.4207355 + 0.4353975x$

f. $P_1(x) = 0.6479184 + 0.5281226x$

5. Los errores de las aproximaciones en el ejercicio 3 son: **a.** 0.177779 **b.** 0.0457206 **c.** 0.00484624 **d.** 0.0526541 **e.** 0.0153784 **f.** 0.00363453

7. El proceso de Gram-Schmidt produce los siguientes grupos de polinomios:

a. $\phi_0(x) = 1,\ \phi_1(x) = x - 0.5,\quad \phi_2(x) = x^2 - x + \frac{1}{6}\quad$ y $\quad \phi_3(x) = x^3 - 1.5x^2 + 0.6x - 0.05$

b. $\phi_0(x) = 1,\ \phi_1(x) = x - 1,\quad \phi_2(x) = x^2 - 2x + \frac{2}{3}\quad$ y $\quad \phi_3(x) = x^3 - 3x^2 + \frac{12}{5}x - \frac{2}{5}$

c. $\phi_0(x) = 1,\ \phi_1(x) = x - 2,\quad \phi_2(x) = x^2 - 4x + \frac{11}{3}\quad$ y $\quad \phi_3(x) = x^3 - 6x^2 + 11.4x - 6.8$

9. Los polinomios de mínimos cuadrados de segundo grado son:

a. $P_2(x) = 3.833333\phi_0(x) + 4\phi_1(x) + 0.9999998\phi_2(x)$

b. $P_2(x) = 2\phi_0(x) + 3.6\phi_1(x) + 3\phi_2(x)$

c. $P_2(x) = 0.5493061\phi_0(x) - 0.2958369\phi_1(x) + 0.1588785\phi_2(x)$

d. $P_2(x) = 3.194528\phi_0(x) + 3\phi_1(x) + 1.458960\phi_2(x)$

e. $P_2(x) = 0.6567600\phi_0(x) + 0.09167105\phi_1(x) - 0.7375118\phi_2(x)$

f. $P_2(x) = 1.471878\phi_0(x) + 1.666667\phi_1(x) + 0.2597705\phi_2(x)$

11. Los polinomios de Laguerre son $L_1(x) = x - 1$, $L_2(x) = x^2 - 4x + 2$ y $L_3(x) = x^3 - 9x^2 + 18x - 6$.

Conjunto de ejercicios 8.3

1. Los polinomios interpolantes de segundo grado son:

a. $P_2(x) = 2.377443 + 1.590534(x - 0.8660254) + 0.5320418(x - 0.8660254)x$

b. $P_2(x) = 0.7617600 + 0.8796047(x - 0.8660254)$

c. $P_2(x) = 1.052926 + 0.4154370(x - 0.8660254) - 0.1384262x(x - 0.8660254)$

d. $P_2(x) = 0.5625 + 0.649519(x - 0.8660254) + 0.75x(x - 0.8660254)$

3. Los polinomios interpolantes de tercer grado son:

a.
$$\begin{aligned} P_3(x) = {} & 2.519044 + 1.945377(x - 0.9238795) \\ & + 0.7047420(x - 0.9238795)(x - 0.3826834) \\ & + 0.1751757(x - 0.9238795)(x - 0.3826834)(x + 0.3826834) \end{aligned}$$

b.
$$\begin{aligned} P_3(x) = {} & 0.7979459 + 0.7844380(x - 0.9238795) - 0.1464394(x - 0.9238795)(x - 0.3826834) \\ & - 0.1585049(x - 0.9238795)(x - 0.3826834)(x + 0.3826834) \end{aligned}$$

c. $P_3(x) = 1.072911 + 0.3782067(x - 0.9238795) - 0.09799213(x - 0.9238795)(x - 0.3826834)$
$+ 0.04909073(x - 0.9238795)(x - 0.3826834)(x + 0.3826834)$

d. $P_3(x) = 0.7285533 + 1.306563(x - 0.9238795) + 0.9999999(x - 0.9238795)(x - 0.3826834)$

5. Los ceros de $\tilde{T}_3$ producen los siguientes polinomios interpolantes de segundo grado.

a. $P_2(x) = 0.3489153 - 0.1744576(x - 2.866025) + 0.1538462(x - 2.866025)(x - 2)$

b. $P_2(x) = 0.1547375 - 0.2461152(x - 1.866025) + 0.1957273(x - 1.866025)(x - 1)$

c. $P_2(x) = 0.6166200 - 0.2370869(x - 0.9330127) - 0.7427732(x - 0.9330127)(x - 0.5)$

d. $P_2(x) = 3.0177125 + 1.883800(x - 2.866025) + 0.2584625(x - 2.866025)(x - 2)$

7. El polinomio cúbico $\frac{383}{384}x - \frac{5}{32}x^3$ aproxima sen x con un error de 7.19×10^{-4}.

9. El cambio de variable $x = \cos\theta$ produce

$$\int_{-1}^{1} \frac{T_n^2(x)}{\sqrt{1-x^2}}\,dx = \int_{-1}^{1} \frac{[\cos(n \arccos x)]^2}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} (\cos(n\theta))^2\,dx = \frac{\pi}{2}.$$

Conjunto de ejercicios 8.4

1. Las aproximaciones de segundo grado de Padé para $f(x) = e^{2x}$ son:

$$n = 2,\ m = 0:\ r_{2,0}(x) = 1 + 2x + 2x^2$$

$$n = 1,\ m = 1:\ r_{1,1}(x) = (1 + x)/(1 - x)$$

$$n = 0,\ m = 2:\ r_{0,2}(x) = (1 - 2x + 2x^2)^{-1}$$

i	x_i	$f(x_i)$	$r_{2,0}(x_i)$	$r_{1,1}(x_i)$	$r_{0,2}(x_i)$
1	0.2	1.4918	1.4800	1.5000	1.4706
2	0.4	2.2255	2.1200	2.3333	1.9231
3	0.6	3.3201	2.9200	4.0000	1.9231
4	0.8	4.9530	3.8800	9.0000	1.4706
5	1.0	7.3891	5.0000	indefinido	1.0000

3. $r_{2,3}(x) = (1 + \frac{2}{5}x + \frac{1}{20}x^2)/(1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3)$

i	x_i	$f(x_i)$	$r_{2,3}(x_i)$
1	0.2	1.22140276	1.22140277
2	0.4	1.49182470	1.49182561
3	0.6	1.82211880	1.82213210
4	0.8	2.22554093	2.22563652
5	1.0	2.71828183	2.71875000

5. $r_{3,3}(x) = (x - \frac{7}{60}x^3)/(1 + \frac{1}{20}x^2)$

i	x_i	$f(x_i)$	Polinomio de MacLaurin de grado 6	$r_{3,3}(x_i)$
0	0.0	0.00000000	0.00000000	0.00000000
1	0.1	0.09983342	0.09966675	0.09938640
2	0.2	0.19866933	0.19733600	0.19709571
3	0.3	0.29552021	0.29102025	0.29246305
4	0.4	0.38941834	0 37875200	0.38483660
5	0.5	0.47942554	0.45859375	0.47357724

7. Las aproximaciones de segundo grado de Padé son:

a. $r_{0,5}(x) = (1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \frac{1}{120}x^5)^{-1}$

b. $r_{1,4}(x) = (1 - \frac{1}{5}x)/(1 + \frac{4}{5}x + \frac{3}{10}x^2 + \frac{1}{15}x^3 + \frac{1}{120}x^4)$

c. $r_{3,2}(x) = (1 - \frac{3}{5}x + \frac{3}{20}x^2 - \frac{1}{60}x^3)/(1 - \frac{2}{5}x + \frac{1}{20}x^2)$

d. $r_{4,1}(x) = (1 - \frac{4}{5}x + \frac{3}{10}x^2 - \frac{1}{15}x^3 + \frac{1}{120}x^4)/(1 + \frac{1}{5}x)$

i	x_i	$f(x_i)$	$r_{0,5}(x_i)$	$r_{1,4}(x_i)$	$r_{2,3}(x_i)$	$r_{4,1}(x_i)$
1	0.2	0.81873075	0.81873081	0.81873074	0.81873075	0.81873077
2	0.4	0.67032005	0.67032276	0.67031942	0.67031963	0.67032099
3	0.6	0.54881164	0.54883296	0.54880635	0.54880763	0.54882143
4	0.8	0.44932896	0.44941181	0.44930678	0.44930966	0.44937931
5	1.0	0.36787944	0.36809816	0.36781609	0.36781609	0.36805556

9. $r_{T_{2,0}}(x) = (1.266066T_0(x) - 1.130318T_1(x) + 0.2714953T_2(x))/T_0(x)$

$r_{T_{1,1}}(x) = (0.9945705T_0(x) - 0.4569046T_1(x))/(T_0(x) + 0.48038745T_1(x))$

$r_{T_{0,2}}(x) = 0.7940220T_0(x)/(T_0(x) + 0.8778575(T_1(x) + 0.1774266T_2(x))$

i	x_i	$f(x_i)$	$r_{T_{2,0}}(x_i)$	$r_{T_{1,1}}(x_i)$	$r_{T_{0,2}}(x_i)$
1	0.25	0.77880078	0.74592811	0.78595377	0.74610974
2	0.50	0.60653066	0.56515935	0.61774075	0.58807059
3	1.00	0.36787944	0.40724330	0.36319269	0.38633199

11. $r_{T_{2,2}}(x) = \dfrac{0.91747T_1(x)}{T_0(x) + 0.088914T_2(x)}$

i	x_i	$f(x_i)$	$r_{T_{2,2}}(x_i)$
0	0.00	0.00000000	0.00000000
1	0.10	0.09983342	0.09093843
2	0.20	0.19866933	0.18028797
3	0.30	0.29552021	0.26808992
4	0.40	0.38941834	0.35438412

13. a. $e^x = e^{M \ln\sqrt{10} + s} = e^{M \ln\sqrt{10}} e^s = e^{\ln 10^{\frac{M}{2}}} e^s = 10^{\frac{M}{2}} e^s$

b. $e^s \approx (1 + \frac{1}{2}s + \frac{1}{10}s^2 + \frac{1}{120}s^3)/(1 - \frac{1}{2}s + \frac{1}{10}s^2 - \frac{1}{120}s^3)$, con $|\text{error}| \leq 3.75 \times 10^{-7}$.

c. Sea M = redondeado $(0.8685889638x)$, $s = x - M/(0.8685889638)$ y
$\hat{f} = (1 + \frac{1}{2}s + \frac{1}{10}s^2 + \frac{1}{120}s^3)/(1 - \frac{1}{2}s + \frac{1}{10}s^2 - \frac{1}{120}s^3)$. Así, $f = (3.16227766)^M \hat{f}$.

Conjunto de ejercicios 8.5

1. $S_2(x) = \frac{\pi^2}{3} - 4\cos x + \cos 2x$

3. $S_3(x) = 3.676078 - 3.676078\cos x + 1.470431\cos 2x - 0.7352156\cos 3x + 3.676078\operatorname{sen} x - 2.940862\operatorname{sen} 2x$

5. $S_n(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{k=1}^{n-1}\frac{1-(-1)^k}{k}\operatorname{sen} kx$

7. Los polinomios trigonométricos de mínimos cuadrados son:

a. $S_2(x) = \cos 2x$

b. $S_2(x) = 0$

c. $S_3(x) = 1.566453 + 0.5886815\cos x - 0.2700642\cos 2x + 0.2175679\cos 3x + 0.8341640\operatorname{sen} x - 0.3097866\operatorname{sen} 2x$

d. $S_3(x) = -2.046326 + 3.883872\cos x - 2.320482\cos 2x + 0.7310818\cos 3x$

9. El polinomio trigonométrico de mínimos cuadrados es $S_3(x) = -0.4968929 + 0.2391965\cos x + 1.515393\cos 2x + 0.2391965\cos 3x - 1.150649\operatorname{sen} x$, con un error de $E(S_3) = 7.271197$.

11. Los polinomios trigonométricos de mínimos cuadrados son:

a. $S_3(x) = -0.08676065 - 1.446416\cos\pi(x-3) - 1.617554\cos 2\pi(x-3) + 3.980729\cos 3\pi(x-3) - 2.154320\operatorname{sen}\pi(x-3) + 3.907451\operatorname{sen} 2\pi(x-3)$, con $E(S_3) = 210.90453$

b. $S_3(x) = -0.0867607 - 1.446416\cos\pi(x-3) - 1.617554\cos 2\pi(x-3) + 3.980729\cos 3\pi(x-3) - 2.354088\cos 4\pi(x-3) - 2.154320\operatorname{sen}\pi(x-3) + 3.907451\operatorname{sen} 2\pi(x-3) - 1.166181\operatorname{sen} 3\pi(x-3)$, con $E(S_4) = 169.4943$

13. Sea $f(-x) = -f(x)$. La integral $\int_{-a}^{0} f(x)\,dx$ con el cambio de variable $t = -x$ se transforma en

$$-\int_{a}^{0} f(-t)\,dt = \int_{0}^{a} f(-t)\,dt = -\int_{0}^{a} f(t)\,dt = -\int_{0}^{a} f(x)\,dx.$$

Por tanto,

$$\int_{-a}^{a} f(x)\,dx = \int_{-a}^{0} f(x)\,dx + \int_{0}^{a} f(x)\,dx = -\int_{0}^{a} f(x)\,dx + \int_{0}^{a} f(x)\,dx = 0.$$

Conjunto de ejercicios 8.6

1. Los polinomios interpolantes trigonométricos son:

a. $S_2(x) = -12.33701 + 4.934802\cos x - 2.467401\cos 2x + 4.934802\operatorname{sen} x$

b. $S_2(x) = -6.168503 + 9.869604\cos x - 3.701102\cos 2x + 4.934802\operatorname{sen} x$

c. $S_2(x) = 1.570796 - 1.570796\cos x$

d. $S_2(x) = -0.5 - 0.5\cos 2x + \operatorname{sen} x$

3. El algoritmo de la transformada rápida de Fourier produce los siguientes polinomios interpolantes trigonométricos.

a. $S_4(x) = -11.10331 + 2.467401\cos x - 2.467401\cos 2x + 2.467401\cos 3x - 1.233701\cos 4x + 5.956833\operatorname{sen} x - 2.467401\operatorname{sen} 2x + 1.022030\operatorname{sen} 3x$

b. $S_4(x) = 1.570796 - 1.340759\cos x - 0.2300378\cos 3x$

c. $S_4(x) = -0.1264264 + 0.2602724\cos x - 0.3011140\cos 2x + 1.121372\cos 3x + 0.04589648\cos 4x - 0.1022190\operatorname{sen} x + 0.2754062\operatorname{sen} 2x - 2.052955\operatorname{sen} 3x$

d. $S_4(x) = -0.1526819 + 0.04754278\cos x + 0.6862114\cos 2x - 1.216913\cos 3x + 1.176143\cos 4x - 0.8179387\operatorname{sen} x + 0.1802450\operatorname{sen} 2x + 0.2753402\operatorname{sen} 3x$

5.	Aproximación	Real
a.	−69.76415	−62.01255
b.	9.869602	9.869604
c.	−0.7943605	−0.2739383
d.	−0.9593287	−0.9557781

7. Todos los términos b_j son cero. Los términos a_j son así:

$a_0 = -4.0008033$	$a_1 = 3.7906715$	$a_2 = -2.2230259$	$a_3 = 0.6258042$
$a_4 = -0.3030271$	$a_5 = 0.1813613$	$a_6 = -0.1216231$	$a_7 = 0.0876136$
$a_8 = -0.0663172$	$a_9 = 0.0520612$	$a_{10} = -0.0420333$	$a_{11} = 0.0347040$
$a_{12} = -0.0291807$	$a_{13} = 0.0249129$	$a_{14} = -0.0215458$	$a_{15} = 0.0188421$
$a_{16} = -0.0166380$	$a_{17} = 0.0148174$	$a_{18} = -0.0132962$	$a_{19} = 0.0120123$
$a_{20} = -0.0109189$	$a_{21} = 0.0099801$	$a_{22} = -0.0091683$	$a_{23} = 0.0084617$
$a_{24} = -0.0078430$	$a_{25} = 0.0072984$	$a_{26} = -0.0068167$	$a_{27} = 0.0063887$
$a_{28} = -0.0060069$	$a_{29} = 0.0056650$	$a_{30} = -0.0053578$	$a_{31} = 0.0050810$
$a_{32} = -0.0048308$	$a_{33} = 0.0046040$	$a_{34} = -0.0043981$	$a_{35} = 0.0042107$
$a_{36} = -0.0040398$	$a_{37} = 0.0038837$	$a_{38} = -0.0037409$	$a_{39} = 0.0036102$
$a_{40} = -0.0034903$	$a_{41} = 0.0033803$	$a_{42} = -0.0032793$	$a_{43} = 0.0031866$
$a_{44} = -0.0031015$	$a_{45} = 0.0030233$	$a_{46} = -0.0029516$	$a_{47} = 0.0028858$
$a_{48} = -0.0028256$	$a_{49} = 0.0027705$	$a_{50} = -0.0027203$	$a_{51} = 0.0026747$
$a_{52} = -0.0026333$	$a_{53} = 0.0025960$	$a_{54} = -0.0025626$	$a_{55} = 0.0025328$
$a_{56} = -0.0025066$	$a_{57} = 0.0024837$	$a_{58} = -0.0024642$	$a_{59} = 0.0024478$
$a_{60} = -0.0024345$	$a_{61} = 0.0024242$	$a_{62} = -0.0024169$	$a_{63} = 0.0024125$

Conjunto de ejercicios 9.1

1. a. Los valores característicos y los vectores característicos asociados son $\lambda_1 = 2$, $\mathbf{v}^{(1)} = (1, 0, 0)^t$; $\lambda_2 = 1$, $\mathbf{v}^{(2)} = (0, 2, 1)^t$ y $\lambda_3 = -1$, $\mathbf{v}^{(3)} = (-1, 1, 1)^t$. Sí, el conjunto es linealmente independiente.

b. Los valores característicos y los vectores característicos asociados son $\lambda_1 = \lambda_2 = \lambda_3 = 1$, $\mathbf{v}^{(1)} = \mathbf{v}^{(2)} = (1, 0, 1)^t$ y $\mathbf{v}^{(3)} = (0, 1, 1)$. El conjunto es linealmente independiente.

c. Los valores característicos y los vectores característicos asociados son $\lambda_1 = 2$, $\mathbf{v}^{(1)} = (0, 1, 0)^t$; $\lambda_2 = 3$, $\mathbf{v}^{(2)} = (1, 0, 1)^t$ y $\lambda_3 = 1$, $\mathbf{v}^{(3)} = (1, 0, -1)^t$. Sí, el conjunto es linealmente independiente.

d. Los valores característicos y los vectores característicos asociados son $\lambda_1 = \lambda_2 = 3$, $\mathbf{v}^{(1)} = (1, 0, -1)^t$, $\mathbf{v}^{(2)} = (0, 1, -1)^t$ y $\lambda_3 = 0$, $\mathbf{v}^{(3)} = (1, 1, 1)^t$. Sí, el conjunto es linealmente independiente.

e. Los valores característicos y los vectores característicos asociados son $\lambda_1 = 1$, $\mathbf{v}^{(1)} = (0, -1, 1)^t$; $\lambda_2 = 1, +\sqrt{2}$, $\mathbf{v}^{(2)} = (\sqrt{2}, 1, 1)^t$ y $\lambda_3 = 1 - \sqrt{2}$, $\mathbf{v}^{(3)} = (-\sqrt{2}, 1, 1)^t$. Sí, el conjunto es linealmente independiente.

f. Los valores característicos y los vectores característicos asociados son $\lambda_1 = 1$, $\mathbf{v}^{(1)} = (1, 0, -1)^t$; $\lambda_2 = 1$, $\mathbf{v}^{(2)} = (1, -1, 0)^t$ y $\lambda_3 = 4$, $\mathbf{v}^{(3)} = (1, 1, 1)^t$. Sí, el conjunto es linealmente independiente.

3. a. Los tres valores característicos se encuentran dentro de $\{\lambda \mid |\lambda| \leq 2\} \cup \{\lambda \mid |\lambda - 2| \leq 2\}$.

b. Los tres valores característicos se encuentran dentro de $R_1 = \{\lambda \mid |\lambda - 4| \leq 2\}$.

c. Los tres valores característicos reales satisfacen $0 \leq \lambda \leq 6$.

d. Los tres valores característicos reales satisfacen $1.25 \leq \lambda \leq 8.25$.

e. Los cuatro valores característicos reales satisfacen $-8 \leq \lambda \leq 1$.

f. Los cuatro valores característicos se encuentran dentro de $R_1 = \{\lambda \mid |\lambda - 2| \leq 4\}$.

5. Si $c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}$, entonces para toda j, con $1 \leq j \leq k$, tenemos $c_1\mathbf{v}_j^t\mathbf{v}_1 + \cdots + c_k\mathbf{v}_j^t\mathbf{v}_k = 0$. Pero la ortogonalidad da $c_i\mathbf{v}_j^t\mathbf{v}_i = 0$, para $i \neq j$, por lo que $c_j\mathbf{v}_j^t\mathbf{v}_j = 0$ y puesto que $\mathbf{v}_j^t\mathbf{v}_j \neq 0$, debemos tener $c_j = 0$.

7. Dado que $\{\mathbf{v}_i\}_{i=1}^{n}$ es linealmente independiente en $\mathbb{R}^n$, existen los números $c_1, \ldots, c_n$ con

$$\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n.$$

Por tanto, para toda k, con $1 \leq k \leq n$,

$$\mathbf{v}_k^t\mathbf{x} = c_1\mathbf{v}_k^t\mathbf{v}_1 + \cdots + c_n\mathbf{v}_k^t\mathbf{v}_n = c_k\mathbf{v}_k^t\mathbf{v}_k = c_k.$$

9. a. Los valores característicos son $\lambda_1 = 5.307857563$, $\lambda_2 = -0.4213112993$, $\lambda_3 = -0.1365462647$ con los vectores característicos asociados $(0.59020967, 0.51643129, 0.62044441)^t$, $(0.77264234, -0.13876278, -0.61949069)^t$ y $(0.23382978, -0.84501102, 0.48091581)^t$, respectivamente.
   b. A no es positiva definida porque $\lambda_2 < 0$ y $\lambda_3 < 0$.

Conjunto de ejercicios 9.2

1. Los valores característicos aproximados y los vectores aproximados son:
   a. $\mu^{(3)} = 3.666667$, $\mathbf{x}^{(3)} = (0.9772727, 0.9318182, 1)^t$
   b. $\mu^{(3)} = 2.000000$, $\mathbf{x}^{(3)} = (1, 1, 0.5)^t$
   c. $\mu^{(3)} = 5.000000$, $\mathbf{x}^{(3)} = (-0.2578947, 1, -0.2842105)^t$
   d. $\mu^{(3)} = 5.038462$, $\mathbf{x}^{(3)} = (1, 0.2213741, 0.3893130, 0.4045802)^t$
   e. $\mu^{(3)} = 7.531073$, $\mathbf{x}^{(3)} = (0.6886722, -0.6706677, -0.9219805, 1)^t$
   f. $\mu^{(3)} = 4.106061$, $\mathbf{x}^{(3)} = (0.1254613, 0.08487085, 0.00922509, 1)^t$
3. Los valores característicos aproximados y los vectores aproximados son:
   a. $\mu^{(3)} = 3.959538$, $\mathbf{x}^{(3)} = (0.5816124, 05545606, 0.5951383)^t$
   b. $\mu^{(3)} = 2.0000000$, $\mathbf{x}^{(3)} = (-0.6666667, -0.6666667, -0.3333333)^t$
   c. $\mu^{(3)} = 7.189567$, $\mathbf{x}^{(3)} = (0.5995308, 0.7367472, 0.3126762)^t$
   d. $\mu^{(3)} = 6.037037$, $\mathbf{x}^{(3)} = (0.5073714, 0.4878571, -0.6634857, -0.2536857)^t$
   e. $\mu^{(3)} = 5.142562$, $\mathbf{x}^{(3)} = (0.8373051, 03701770, 0.1939022, 0.3525495)^t$
   f. $\mu^{(3)} = 8.593142$, $\mathbf{x}^{(3)} = (-0.4134762, 0.4026664, 0.5535536, -0.6003962)^t$
5. Los valores característicos aproximados y los vectores aproximados son:
   a. $\lambda_1 \approx \mu^{(9)} = 3.999908$, $\mathbf{x}^{(9)} = (0.9999943, 0.9999828, 1)^t$
      $\lambda_2 \approx \mu^{(1)} = 1.000000$, $\mathbf{x}^{(1)} = (-2.999908, 2.999908, 0)^t$
   b. $\lambda_1 \approx \mu^{(13)} = 2.414214$, $\mathbf{x}^{(13)} = (1, 0.7071429, 0.7070707)^t$
      $\lambda_2 \approx \mu^{(1)} = 1.000000$, $\mathbf{x}^{(1)} = (0, -1.414214, 1.414214)^t$
   c. $\lambda_1 \approx \mu^{(9)} = 5.124749$, $\mathbf{x}^{(9)} = (-0.2424476, 1, -0.3199733)^t$
      $\lambda_2 \approx \mu^{(6)} = 1.636734$, $\mathbf{x}^{(6)} = (1.783218, -1.135350, -3.124733)^t$
   d. $\lambda_1 \approx \mu^{(24)} = 5.235861$, $\mathbf{x}^{(24)} = (1, 0.6178361, 0.1181667, 0.4999220)^t$
      $\lambda_2 \approx \mu^{(10)} = 3.618177$, $\mathbf{x}^{(10)} = (0.7236390, -1.170573, 1.170675, -0.2763374)^t$
   e. $\lambda_1 \approx \mu^{(17)} = 8.999667$, $\mathbf{x}^{(17)} = (0.9999085, -0.9999078, -0.9999993, 1)^t$
      $\lambda_2 \approx \mu^{(12)} = 5.000051$, $\mathbf{x}^{(21)} = (1.999338, -1.999603, 1.999603, -2.000198)^t$
   f. El método no convergió en 25 iteraciones, pero $\lambda_1 \approx \mu^{(363)} = 4.105309$, $\mathbf{x}^{(363)} = (0.06286299, 0.08702754, 0.01824680, 1)^t$, $\lambda_2 \approx \mu^{(15)} = -4.024308$, $\mathbf{x}^{(15)} = (-8.151965, 2.100699, 0.7519080, -0.3554941)^t$.
7. Los valores característicos aproximados y los vectores característicos aproximados son:
   a. $\mu^{(8)} = 4.0000000$, $\mathbf{x}^{(8)} = (0.5773547, 0.5773282, 0.5773679)^t$
   b. $\mu^{(13)} = 2.414214$, $\mathbf{x}^{(13)} = (-0.7071068, 0.5000255, -0.4999745)^t$

c. $\mu^{(16)} = 7.223663$, $\mathbf{x}^{(16)} = (0.6247845, 0.7204271, 0.3010466)^t$

d. $\mu^{(20)} = 7.086130$, $\mathbf{x}^{(20)} = (0.3325999, 0.2671862, -0.7590108, -0.4918246)^t$

e. $\mu^{(21)} = 5.236068$, $\mathbf{x}^{(21)} = (0.7795539, 0.4815996, 0.09214214, 0.3897016)^t$

f. $\mu^{(16)} = 9.0000000$, $\mathbf{x}^{(16)} = (-0.4999592, 0.4999584, 0.5000408, -0.5000416)^t$

9. Los valores característicos aproximados y los vectores característicos aproximados son:

a. $\mu^{(9)} = 1.000000$, $\mathbf{x}^{(9)} = (-0.1542994, 0.7715207, -0.6172095)^t$

b. $\mu^{(12)} = -0.4142136$, $\mathbf{x}^{(12)} = (-0.7071068, 0.4999894, 0.5000106)^t$

c. $\mu^{(6)} = 4.961699$, $\mathbf{x}^{(6)} = (-0.4812465, 0.05195336, 0.8750444)^t$

d. $\mu^{(14)} = 2.485863$, $\mathbf{x}^{(14)} = (-0.6096695, 0.6451951, -0.2779286, 0.3671268)^t$

e. $\mu^{(10)} = 3.618034$, $\mathbf{x}^{(10)} = (0.3958550, -0.6404796, 0.6404886, -0.1511924)^t$

f. $\mu^{(6)} = 4.0000000$, $\mathbf{x}^{(6)} = (-0.4999985, -0.5000015, -0.4999985, -0.5000015)^t$

11. Los valores característicos aproximados y los vectores característicos aproximados son:

a. $\mu^{(2)} = 1.000000$, $\mathbf{x}^{(2)} = (0.1542373, -0.7715828, 0.6171474)^t$

b. $\mu^{(13)} = 1.000000$, $\mathbf{x}^{(13)} = (0.00007432, -0.7070723, 0.7071413)^t$

c. $\mu^{(14)} = 4.961699$, $\mathbf{x}^{(14)} = (-0.4814472, 0.05180473, 0.8749428)^t$

d. $\mu^{(17)} = 4.428007$, $\mathbf{x}^{(17)} = (0.7194230, 0.4231908, 0.1153589, 0.5385466)^t$

e. $\mu^{(10)} = 3.618034$, $\mathbf{x}^{(10)} = (0.3956185, -0.6406258, 0.6404462, -0.1513711)^t$

f. El método no convergió en 25 iteraciones, pero $\mu^{(31)} = 5.0000000$, $\mathbf{x}^{(31)} = (0.4999091, -0.5002392, 0.4997607, -0.50009009)^t$.

13. a. Tenemos $|\lambda| \leq 6$ para todos los valores característicos λ.

b. El valor característico aproximado es $\mu^{(133)} = 0.69766854$, con el vector característico aproximado $\mathbf{x}^{(133)} = (1, 0.7166727, 0.2568099, 0.04601217)^t$.

d. El polinomio característico es $P(\lambda) = \lambda^4 - \frac{1}{4}\lambda - \frac{1}{16}$ y los valores característicos son $\lambda_1 = 0.6976684972$, $\lambda_2 = -0.2301775942 + 0.56965884i$, $\lambda_3 = -0.2301775942 - 0.56965884i$ y $\lambda_4 = -0.237313308$.

e. La población de escarabajos debería acercarse a cero, porque A es convergente.

15. Al aplicar el método de potencias inversas con $\mathbf{x}^{(0)} = (1, 0, 0, 1, 0, 0, 1, 0, 0, 1)^t$ y con $q = 0$, se obtienen los siguientes resultados:

a. $\mu^{(49)} = 1.0201926$, así $\rho(A^{-1}) \approx 1/\mu^{(49)} = 0.9802071$;

b. $\mu^{(30)} = 1.0404568$, así $\rho(A^{-1}) \approx 1/\mu^{(30)} = 0.9611163$;

c. $\mu^{(22)} = 1.0606974$, así $\rho(A^{-1}) \approx 1/\mu^{(22)} = 0.9427760$.

El método parece ser estable para toda α en $[\frac{1}{4}, \frac{3}{4}]$.

17. Al formar $A^{-1}B$ y aplicar el método de potencias con $\mathbf{x}^{(0)} = (1, 0, 0, 1, 0, 0, 1, 0, 0, 1)^t$ se obtienen los siguientes resultados:

a. El radio espectral es aproximadamente $\mu^{(46)} = 0.9800021$.

b. El radio espectral es aproximadamente $\mu^{(25)} = 0.9603543$.

c. El radio espectral es aproximadamente $\mu^{(18)} = 0.9410754$.

Conjunto de ejercicios 9.3

1. El método de Householder produce las siguientes matrices tridiagonales.

a. $\begin{bmatrix} 12.00000 & -10.77033 & 0.0 \\ -10.77033 & 3.862069 & 5.344828 \\ 0.0 & 5.344828 & 7.137931 \end{bmatrix}$ **b.** $\begin{bmatrix} 2.0000000 & 1.414214 & 0.0 \\ 1.414214 & 1.000000 & 0.0 \\ 0.0 & 0.0 & 3.0 \end{bmatrix}$

c. $\begin{bmatrix} 1.0000000 & -1.414214 & 0.0 \\ -1.414214 & 1.000000 & 0.0 \\ 0.0 & 0.0 & 1.000000 \end{bmatrix}$ **d.** $\begin{bmatrix} 4.750000 & -2.263846 & 0.0 \\ -2.263846 & 4.475610 & -1.219512 \\ 0.0 & -1.219512 & 5.024390 \end{bmatrix}$

3. El método de Householder produce las siguientes matrices tridiagonales.

a. $\begin{bmatrix} 2.0000000 & 2.8284271 & 1.4142136 \\ -2.8284271 & 1.0000000 & 2.0000000 \\ 0.0000000 & 2.0000000 & 3.0000000 \end{bmatrix}$ **b.** $\begin{bmatrix} -1.0000000 & -3.0655513 & 0.0000000 \\ -3.6055513 & -0.23076923 & 3.1538462 \\ 0.0000000 & 0.15384615 & 2.2307692 \end{bmatrix}$

c. $\begin{bmatrix} 5.0000000 & 4.9497475 & -1.4320780 & -1.5649769 \\ -1.4142136 & -2.0000000 & -2.4855515 & 1.8226448 \\ 0.0000000 & -5.4313902 & -1.4237288 & -2.6486542 \\ 0.0000000 & 0.0000000 & 1.5939865 & 5.4237288 \end{bmatrix}$

d. $\begin{bmatrix} 4.0000000 & 1.7320508 & 0.0000000 & 0.0000000 \\ 1.7320508 & 2.3333333 & 0.23570226 & 0.40824829 \\ 0.0000000 & -0.47140452 & 4.6666667 & -0.57735027 \\ 0.0000000 & 0.0000000 & 0.0000000 & 5.0000000 \end{bmatrix}$

Conjunto de ejercicios 9.4

1. Dos iteraciones del algoritmo QR producen las siguientes matrices.

a. $A^{(3)} = \begin{bmatrix} 0.6939977 & -0.3759745 & 0.0 \\ -0.3759745 & 1.892417 & -0.03039696 \\ 0.0 & -0.03039696 & 3.413585 \end{bmatrix}$

b. $A^{(3)} = \begin{bmatrix} 4.535466 & 1.212648 & 0.0 \\ 1.212648 & 3.533242 & 3.83 \times 10^{-7} \\ 0.0 & 3.83 \times 10^{-7} & -0.06870782 \end{bmatrix}$

c. $A^{(3)} = \begin{bmatrix} 4.679567 & -0.2969009 & 0.0 \\ -2.969009 & 3.052484 & -1.207346 \times 10^{-5} \\ 0.0 & -1.207346 \times 10^{-5} & 1.267949 \end{bmatrix}$

d. $A^{(3)} = \begin{bmatrix} 0.3862092 & 0.4423226 & 0.0 & 0.0 \\ 0.4423226 & 1.787694 & -0.3567744 & 0.0 \\ 0.0 & -0.3567744 & 3.080815 & 3.116382 \times 10^{-5} \\ 0.0 & 0.0 & 3.116382 \times 10^{-5} & 4.745281 \end{bmatrix}$

e. $A^{(3)} = \begin{bmatrix} -2.826365 & 1.130297 & 0.0 & 0.0 \\ 1.130297 & -2.429647 & -0.1734156 & 0.0 \\ 0.0 & -0.1734156 & 0.8172086 & 1.863997 \times 10^{-9} \\ 0.0 & 0.0 & 1.863997 \times 10^{-9} & 3.438803 \end{bmatrix}$

f. $A^{(3)} = \begin{bmatrix} 0.2763388 & 0.1454371 & 0.0 & 0.0 \\ 0.1454371 & 0.4543713 & 0.1020836 & 0.0 \\ 0.0 & 0.1020836 & 1.174446 & -4.36 \times 10^{-5} \\ 0.0 & 0.0 & -4.36 \times 10^{-5} & 0.9948441 \end{bmatrix}$

3. Las matrices del ejercicio 1 tienen los siguientes valores característicos, con una exactitud de 10^{-5}.

a. 3.414214, 2.000000, 0.58578644

b. −0.06870782, 5.346462, 2.722246

c. 1.267949, 4.732051, 3.000000

d. 4.745281, 3.177283, 1.822717, 0.2547188

e. 3.438803, 0.8275517, −1.488068, −3.778287

f. 0.9948440, 1.189091, 0.5238224, 0.1922421

5. a. Sea

$$P = \begin{bmatrix} \cos\theta & -\text{sen}\theta \\ \text{sen}\theta & \cos\theta \end{bmatrix}$$

y $\mathbf{y} = P\mathbf{x}$. Demuestre que $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$. Use la relación $x_1 + ix_2 = re^{i\alpha}$, en la que $r = \|\mathbf{x}\|_2$ y $\alpha = \tan^{-1}(x_2/x_1)$ y $y_1 + iy_2 = re^{i(\alpha+\theta)}$.

b. Sea $\mathbf{x} = (1, 0)^t$ y $\theta = \pi/4$.

11. a. Con una exactitud de 10^{-5}, los valores característicos son 2.618034, 3.618034, 1.381966 y 0.3819660.

b. En función de p y ρ, los valores característicos son $-65.45085p/\rho$, $-90.45085p/\rho$, $-34.54915p/\rho$ y $-9.549150p/\rho$.

13. Los valores característicos reales son:

a. Cuando $\alpha = 1/4$, tenemos 0.97974649, 0.92062677, 0.82743037, 0.70770751, 0.57115742, 0.42884258, 0.29229249, 0.17256963, 07937323 y 0.02025351.

b. Cuando $\alpha = 1/2$, tenemos 0.95949297, 0.84125353, 0.65486073, 0.41541501, 0.14231484, −0.14231484, −0.41541501, −0.65486073, −0.84125353 y −0.95949297.

c. Cuando $\alpha = 3/4$, tenemos 0.93923946, 0.76188030, 0.48229110, 0.12312252, −0.28652774, −0.71347226, −1.13212252, −1.48229110, −1.76188030 y −1.93923946. El método parece ser estable para $\alpha \leq \frac{1}{2}$.

Conjunto de ejercicios 10.1

1. Use el teorema 10.5.

3. Use el teorema 10.5 con cada una de las derivadas parciales.

5. b. Con $\mathbf{x}^{(0)} = (0, 0)^t$ y tolerancia 10^{-5}, tenemos $\mathbf{x}^{(13)} = (0.9999973, 0.9999973)^t$.

c. Con $\mathbf{x}^{(0)} = (0, 0)^t$ y tolerancia 10^{-5}, tenemos $\mathbf{x}^{(11)} = (0.9999984, 0.9999991)^t$.

7. a. Con $\mathbf{x}^{(0)} = (1, 1, 1)^t$ tenemos $\mathbf{x}^{(5)} = (5{,}0000000, 0.0000000, -0.5235988)^t$.

b. Con $\mathbf{x}^{(0)} = (1, 1, 1)^t$ tenemos $\mathbf{x}^{(9)} = (1.0364011, 1.0857072, 0.93119113)^t$.

c. Con $\mathbf{x}^{(0)} = (0, 0, 0.5)^t$ tenemos $\mathbf{x}^{(5)} = (0.00000000, 0.09999999, 1{,}0000000)^t$.

d. Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$ tenemos $\mathbf{x}^{(5)} = (0.49814471, -0.19960600, -0.52882595)^t$.

9. a. Con $\mathbf{G}(\mathbf{x}) = \left(\sqrt{x_1 - x_2^2}, \sqrt{x_1^2 - x_2}\right)^t$ y $\mathbf{x}^{(0)} = (0.7, 0.4)^t$, tenemos $\mathbf{x}^{(14)} = (0.77184647, 0.41965131)^t$.

b. Con $\mathbf{G}(\mathbf{x}) = \left(x/\sqrt{3}, \sqrt{(1 + x_1^3)/(3x_1)}\right)^t$ y $\mathbf{x}^{(0)} = (0.4, 0.7)^t$, tenemos $\mathbf{x}^{(20)} = (0.4999980, 0.8660221)^t$.

c. Con $\mathbf{G}(\mathbf{x}) = (\sqrt{37 - x_2}, \sqrt{x_1 - 5}, 3 - x_1 - x_2)^t$ y $\mathbf{x}^{(0)} = (5, 1, -1)^t$, tenemos $\mathbf{x}^{(10)} = (6.0000002, 1.0000000, -3.9999971)^t$.

d. Con $\mathbf{G}(\mathbf{x}) = \left(\sqrt{2x_3 + x_2 - 2x_2^2}, \sqrt{(10x_3 + x_1^2)/8}, x_1^2/(7x_2)\right)^t$ y $\mathbf{x}^{(0)} = (0.5, 0.5, 0)^t$, tenemos $\mathbf{x}^{(60)} = (0.5291548, 0.4000018, 0.09999853)^t$.

11. Sí, ocurre una solución estable cuando $x_1 = 8000$ y $x_2 = 4000$.

Conjunto de ejercicios 10.2

1. a. $\mathbf{x}^{(2)} = (0.4958936, 1.983423)^t$

b. $\mathbf{x}^{(2)} = (-0.5131616, -0.01837622)^t$

c. $\mathbf{x}^{(2)} = (0.5001667, 0.2508036, -0.5173874)^t$

d. $\mathbf{x}^{(2)} = (4.350877, 18.49123, -19.84211)^t$

3. a. $\mathbf{x}^{(5)} = (0.5000000, 0.8660254)^t$

b. $\mathbf{x}^{(6)} = (1.772454, 1.772454)^t$

c. $\mathbf{x}^{(5)} = (-1.456043, -1.664230, 0.4224934)^t$

d. $\mathbf{x}^{(4)} = (0.4981447, -0.1996059, -0.5288260)^t$

5. Con $\mathbf{x}^{(0)} = (1, 1 - 1)^t$ y $TOL = 10^{-6}$, tenemos $\mathbf{x}^{(20)} = (0.5, 9.5 \times 10^{-7}, -0.5235988)^t$.

7. Cuando la dimensión n es 1, $\mathbf{F}(\mathbf{x})$ es una función de un componente $f(\mathbf{x}) = f_1(\mathbf{x})$, y el vector $\mathbf{x}$ tiene sólo un componente $x_1 = x$. En este caso, la matriz jacobiana $J(\mathbf{x})$ se reduce a la matriz 1×1 $\left[\frac{\partial f_1}{\partial x_1}(\mathbf{x})\right] = f'(\mathbf{x}) = f'(x)$.
Por tanto, la ecuación de vector

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - J(\mathbf{x}^{(k-1)})^{-1}\mathbf{F}(\mathbf{x}^{(k-1)})$$

se convierte en la ecuación escalar

$$x_k = x_{k-1} - f(x_{k-1})^{-1} f(x_{k-1}) = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}.$$

9. Con $\theta_i^{(0)} = 1$, para toda $i = 1, 2, \ldots, 20$ se obtienen los siguientes resultados:

i	1	2	3	4	5	6
$\theta_i^{(5)}$	0.14062	0.19954	0.24522	0.28413	0.31878	0.35045

i	7	8	9	10	11	12	13
$\theta_i^{(5)}$	0.37990	0.40763	0.43398	0.45920	0.48348	0.50697	0.52980

i	14	15	16	17	18	19	20
$\theta_i^{(5)}$	0.55205	0.57382	0.59516	0.61615	0.63683	0.65726	0.67746

11. **a.** Tenemos

$$\frac{\partial E}{\partial a} = 2\sum_{i=1}^{n}\left(w_i y_i - \frac{a}{(x_i - b)^c}\right)\left(\frac{1}{(x_i - b)^c}\right) = 0,$$
$$\frac{\partial E}{\partial b} = 2\sum_{i=1}^{n}\left(w_i y_i - \frac{a}{(x_i - b)^c}\right)\left(\frac{-ac}{(x_i - b)^{c+1}}\right) = 0,$$

y

$$\frac{\partial E}{\partial c} = 2\sum_{i=1}^{n}\left(w_i y_i - \frac{a}{(x_i - b)^c}\right)\ln(x_i - b)\left(\frac{-a}{(x_i - b)^c}\right) = 0.$$

Al resolver a en la primera ecuación y al sustituir en la segunda y tercera ecuaciones, se obtiene el sistema lineal.

b. Con $\mathbf{x}^{(0)} = (26, 8, 8.3)^t = (b_0, c_0)^t$, tenemos $\mathbf{x}^{(7)} = (26.77021, 8.451831)^t$. Por tanto, $a = 2.217952 \times 10^6$, $b = 26.77021$, $c = 8.451831$ y

$$\sum_{i=1}^{n}\left(w_i y_i - \frac{a}{(x_i - b)^c}\right)^2 = 0.7821139.$$

Conjunto de ejercicios 10.3

1. a. $\mathbf{x}^{(2)} = (0.4777920, 1.927557)^t$

b. $\mathbf{x}^{(2)} = (-0.3250070, -0.1386967)^t$

c. $\mathbf{x}^{(2)} = (0.5115893, -78.72872, -0.5120771)^t$

d. $\mathbf{x}^{(2)} = (-67.00583, 35.06480, -123.3408)^t$

3. a. $\mathbf{x}^{(9)} = (0.5, 0.8660254)^t$

b. $\mathbf{x}^{(8)} = (1.772454, 1.772454)^t$

c. $\mathbf{x}^{(9)} = (-1.456043, -1.664231, -0.4224934)^t$

d. $\mathbf{x}^{(5)} = (0.4981447, -0.1996059, -0.5288260)^t$

5. Con $\mathbf{x}^{(0)} = (1, 1 - 1)^t$, tenemos $\mathbf{x}^{(56)} = (0.5000591, 0.01057235, -0.5224818)^t$.

7. Con $\mathbf{x}^{(0)} = (0.75, 1.25)^t$, tenemos $\mathbf{x}^{(4)} = (0.7501948, 1.184712)^t$. Por tanto, $a = 0.7501948$, $b = 1.184712$ y el error es 19.796.

Conjunto de ejercicios 10.4

1. a. Con $\mathbf{x}^{(0)} = (0, 0)^t$, tenemos $\mathbf{x}^{(11)} = (0.4943541, 1.948040)^t$.

b. Con $\mathbf{x}^{(0)} = (1, 1)^t$, tenemos $\mathbf{x}^{(2)} = (0.4970073, 0.8644143)^t$.

c. Con $\mathbf{x}^{(0)} = (2, 2)^t$, tenemos $\mathbf{x}^{(1)} = (1.736083, 1.804428)^t$.

d. Con $\mathbf{x}^{(0)} = (0, 0)^t$, tenemos $\mathbf{x}^{(2)} = (-0.3610092, 0.05788368)^t$.

3. a. Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$, tenemos $\mathbf{x}^{(14)} = (1.043605, 1.064058, 0.9246118)^t$.

b. Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$, tenemos $\mathbf{x}^{(9)} = (0.4932739, 0.9863888, -0.5175964)^t$.

c. Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$, tenemos $\mathbf{x}^{(11)} = (-1.608296, -1.192750, 0.7205642)^t$.

d. Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$, tenemos $\mathbf{x}^{(1)} = (0, 0.00989056, 0.9890556)^t$.

5. a. Con $\mathbf{x}^{(0)} = (0, 0)^t$, tenemos $\mathbf{x}^{(8)} = (3.136548, 0)^t$ y $g(\mathbf{x}^{(8)}) = 0.005057848$.

b. Con $\mathbf{x}^{(0)} = (0, 0)^t$, tenemos $\mathbf{x}^{(13)} = (0.6157412, 0.3768953)^t$ y $g(\mathbf{x}^{(13)}) = 0.1481574$.

c. Con $\mathbf{x}^{(0)} = (0, 0, 0)^t$, tenemos $\mathbf{x}^{(5)} = (-0.6633785, 0.3145720, 0.5000740)^t$ y $g(\mathbf{x}^{(5)}) = 0.6921548$.

d. Con $\mathbf{x}^{(0)} = (1, 1, 1)^t$, tenemos $\mathbf{x}^{(4)} = (0.04022273, 0.01592477, 0.01594401)^t$ y $g(\mathbf{x}^{(4)}) = 1.010003$.

Conjunto de ejercicios 10.5

1. a. $(3, -2.25)^t$

b. $(0.42105263, 2.6184211)^t$

c. $(2.173110, -1.3627731)^t$

3. Con $\mathbf{x}(0) = \mathbf{0}$ en todas las partes, se obtiene:

a. $(0.44006047, 1.8279835)^t$

b. $(-0.41342613, 0.096669468)^t$

c. $(0.49858909, 0.24999091, -0.52067978)^t$

d. $(6.1935484, 18.532258, -21.725806)^t$

5. a. Con $\mathbf{x}(0) = (-1, 3.5)^t$ da $(-1, 3.5)^t$.

Con $\mathbf{x}(0) = (2.5, 4.0)^t$ da $(2.5469465, 3.9849975)^t$.

b. Con $\mathbf{x}(0) = (0.11, 0.27)^t$ da $(0.12124195, 0.27110516)^t$.

c. Con $\mathbf{x}(0) = (1, 1, 1)^t$ da $(1.0364005, 1.0857066, 0.93119144)^t$.

d. Con $\mathbf{x}(0) = (1, -1, 1)^t$ da $(0.90016074, -1.0023801, 0.49661093)^t$.

Con $\mathbf{x}(0) = (1, 1, -1)^t$ da $(0.50104035, 1.0023801, -0.49661093)^t$.

7. a. $(0.49998949, 0.86608576)^t$
   b. $(1.7724820, 1.7722940)^t$
   c. $(-1.4561027, -1.6642463, 0.42241506)^t$
   d. $(0.49814392, -0.19960453, -0.52882611)^t$
9. $(0.50024553, 0.078230039, -0.52156996)^t$
11. Por cada λ, tenemos

$$0 = G(\lambda, \mathbf{x}(\lambda)) = F(\mathbf{x}(\lambda)) - e^{-\lambda}F(\mathbf{x}(0)),$$

por lo que

$$0 = \frac{\partial F(\mathbf{x}(\lambda))}{\partial \mathbf{x}}\frac{d\mathbf{x}}{d\lambda} + e^{-\lambda} F(\mathbf{x}(0)) = J(\mathbf{x}(\lambda))\mathbf{x}'(\lambda) + e^{-\lambda} F(\mathbf{x}(0))$$

y

$$J(\mathbf{x}(\lambda))x'(\lambda) = -e^{-\lambda} F(\mathbf{x}(0)) = -F(\mathbf{x}(0)).$$

Por consiguiente,

$$x'(\lambda) = -J(\mathbf{x}(\lambda))^{-1} F(\mathbf{x}(0)).$$

Con $N = 1$ tenemos $h = 1$, por lo que

$$\mathbf{x}(1) = \mathbf{x}(0) - J(\mathbf{x}(0))^{-1} F(\mathbf{x}(0)).$$

Sin embargo, el método de Newton da

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - J(\mathbf{x}^{(0)})^{-1}F(\mathbf{x}^{(0)}).$$

Puesto que $\mathbf{x}(0) = \mathbf{x}^{(0)}$, tenemos $\mathbf{x}(1) = \mathbf{x}^{(1)}$.

Conjunto de ejercicios 11.1

1. El algoritmo del disparo lineal produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	w_{1i}	$y(x_i)$
1	0.5	0.82432432	0.82402714

b.

i	x_i	w_{1i}	$y(x_i)$
1	0.25	0.3937095	0.3936767
2	0.50	0.8240948	0.8240271
3	0.75	1.337160	1.337086

3. El algoritmo del disparo lineal produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	w_{1i}	$y(x_i)$
3	0.3	0.7833204	0.7831923
6	0.6	0.6023521	0.6022801
9	0.9	0.8568906	0.8568760

b.

i	x_i	w_{1i}	$y(x_i)$
5	1.25	0.1676179	0.1676243
10	1.50	0.4581901	0.4581935
15	1.75	0.6077718	0.6077740

c.

i	x_i	w_{1i}	$y(x_i)$
3	0.3	−0.5185754	−0.5185728
6	0.6	−0.2195271	−0.2195247
9	0.9	−0.0406577	−0.0406570

d.

i	x_i	w_{1i}	$y(x_i)$
3	1.3	0.0655336	0.06553420
6	1.6	0.0774590	0.07745947
9	1.9	0.0305619	0.03056208

5. El algoritmo del disparo lineal con $h = 0.05$ produce los siguientes resultados.

i	x_i	w_{1i}
6	0.3	0.04990547
10	0.5	0.00673795
16	0.8	0.00033755

El algoritmo del disparo lineal con $h = 0.1$ produce los siguientes resultados.

i	x_i	w_{1i}
3	0.3	0.05273437
5	0.5	0.00741571
8	0.8	0.00038976

7. a. El potencial aproximado es $u(3) \approx 36.66702$ con $h = 0.1$.

b. El potencial real es $u(3) = 36.66667$.

9. a. No hay soluciones si b es un múltiplo entero de π y $B \neq 0$.

b. Existe una solución única siempre que b no sea un múltiplo entero de π.

c. Existe un número de soluciones si b es un múltiplo entero de π y si $B = 0$.

Conjunto de ejercicios 11.2

1. El algoritmo del disparo no lineal produce $w_1 = 0.405505 \approx \ln 1.5 = 0.405465$.

3. El algoritmo del disparo no lineal produce los resultados que se incluyen en las tablas anexas.

a. Se requieren 4 iteraciones:

i	x_i	w_{1i}	$y(x_i)$
3	1.3	0.4347934	0.4347826
6	1.6	0.3846363	0.3846154
9	1.9	0.3448586	0.3448276

b. Se requieren 6 iteraciones:

i	x_i	w_{1i}	$y(x_i)$
3	1.3	2.069249	2.069231
6	1.6	2.225013	2.225000
9	1.9	2.426317	2.426316

c. Se requieren 3 iteraciones:

i	x_i	w_{1i}	$y(x_i)$
3	2.3	1.2676912	1.2676917
6	2.6	1.3401256	1.3401268
9	2.9	1.4095359	1.4095383

d. Para aplicar el algoritmo necesitamos redefinir que el valor inicial de TK sea 2.

Se requieren 7 iteraciones:

i	x_i	w_{1i}	$y(x_i)$
5	1.25	0.4358290	0.4358272
10	1.50	1.3684496	1.3684447
15	1.75	2.9992010	2.9991909

5. El algoritmo produce los resultados que se incluyen en las tablas anexas.

a. Se requieren 3 iteraciones:

i	x_i	w_{1i}	$y(x_i)$
3	1.3	0.4347720	0.4347826
6	1.6	0.3845947	0.3846154
9	1.9	0.3447969	0.3448276

b. Para aplicar el algoritmo necesitamos definir que las aproximaciones iniciales de t_k sean -0.5 y 0.5. Se requieren 15 iteraciones:

i	x_i	w_{1i}	$y(x_i)$
3	1.3	2.0692491	2.0692308
6	1.6	2.2250137	2.2250000
9	1.9	2.4263174	2.4263158

Conjunto de ejercicios 11.3

1. El algoritmo lineal de diferencias finitas produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	w_{1i}	$y(x_i)$
1	0.5	0.83333333	0.82402714

b.

i	x_i	w_{1i}	$y(x_i)$
1	0.25	0.39512472	0.39367669
2	0.5	0.82653061	0.82402714
3	0.75	1.33956916	1.33708613

c. $\dfrac{4(0.82653061) - 0.83333333}{3} = 0.82426304$

3. El algoritmo lineal de diferencias finitas produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	w_i	$y(x_i)$
2	0.2	1.018096	1.0221404
5	0.5	0.5942743	0.59713617
7	0.7	0.6514520	0.65290384

b.

i	x_i	w_i	$y(x_i)$
5	1.25	0.16797186	0.16762427
10	1.50	0.45842388	0.45819349
15	1.75	0.60787334	0.60777401

c.

i	x_i	w_i	$y(x_i)$
3	0.3	−0.5183084	−0.5185728
6	0.6	−0.2192657	−0.2195247
9	0.9	−0.0405748	−0.04065697

c.

i	x_i	w_i	$y(x_i)$
3	1.3	0.0654387	0.0655342
6	1.6	0.0773936	0.0774595
9	1.9	0.0305465	0.0305621

5. El algoritmo lineal de diferencias finitas produce los resultados que se incluyen en las tablas anexas.

i	x_i	$w_i(h = 0.1)$
3	0.3	0.05572807
6	0.6	0.00310518
9	0.9	0.00016516

i	x_i	$w_i(h = 0.05)$
6	0.3	0.05132396
12	0.6	0.00263406
18	0.9	0.00013340

7. a. Las deflexiones aproximadas se incluyen en la tabla anexa.

i	x_i	w_{1i}
5	30	0.0102808
10	60	0.0144277
15	90	0.0102808

b. Sí.

c. Sí; la deflexión ocurre en $x = 60$. La solución exacta se encuentra dentro de la tolerancia, no así la aproximación.

Conjunto de ejercicios 11.4

1. El algoritmo no lineal de diferencias finitas produce los siguientes resultados.

i	x_i	w_i	$y(x_i)$
1	1.5	0.4067967	0.4054651

3. El algoritmo no lineal de diferencias finitas produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	w_{1i}	$y(x_i)$
3	1.3	0.4347972	0.4347826
6	1.6	0.3846286	0.3846154
9	1.9	0.3448316	0.3448276

b.

i	x_i	w_{1i}	$y(x_i)$
3	1.3	2.0694081	2.0692308
6	1.6	2.2250937	2.2250000
9	1.9	2.4263387	2.4263158

c.

i	x_i	w_{1i}	$y(x_i)$
3	2.3	1.2677078	1.2676917
6	2.6	1.3401418	1.3401268
9	2.9	1.4095432	1.4095383

d.

i	x_i	w_{1i}	$y(x_i)$
5	1.25	0.4345979	0.4358273
10	1.50	1.3662119	1.3684447
15	1.75	2.9969339	2.9991909

Conjunto de ejercicios 11.5

1. El algoritmo lineal segmentario produce $\phi(x) = 0.07713274\phi_1(x) - 0.07442678\phi_2(x)$. Los valores reales son $y(x_1) = -0.07988545$ y $y(x_2) = -0.07712903$.

3. El algoritmo segmentario lineal produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	$\phi(x_i)$	$y(x_i)$
3	0.3	−0.212333	−0.21
6	0.6	−0.241333	−0.24
9	0 9	−0.090333	−0.09

b.

i	x_i	$\phi(x_i)$	$y(x_i)$
3	0.3	0.1815138	0.1814273
6	0.6	0.1805502	0.1804753
9	0.9	0.05936468	0.05934303

c.

i	x_i	$\phi(x_i)$	$y(x_i)$
5	0.25	−0.3585989	−0.3585641
10	0.50	−0.5348383	−0.5347803
15	0.75	−0.4510165	−0.4509614

d.

i	x_i	$\phi(x_i)$	$y(x_i)$
5	0.25	−0.1846134	−0.1845204
10	0.50	−0.2737099	−0.2735857
15	0.75	−0.2285169	−0.2284204

5. El algoritmo del trazador cúbico produce los resultados que se incluyen en las tablas anexas.

a.

i	x_i	$\phi(x_i)$	$y(x_i)$
3	0.3	−0.2100000	−0.21
6	0.6	−0.2400000	−0.24
9	0.9	−0.0900000	−0.09

b.

i	x_i	$\phi(x_i)$	$y(x_i)$
3	0.3	0.1814269	0.1814273
6	0.6	0.1804753	0.1804754
9	0.9	0.05934321	0.05934303

c.

i	x_i	$\phi(x_i)$	$y(x_i)$
5	0.25	−0.3585639	−0.3585641
10	0.50	−0.5347779	−0.5347803
15	0.75	−0.4509109	−0.4509614

c.

i	x_i	$\phi(x_i)$	$y(x_i)$
5	0.25	−0.1845191	−0.1845204
10	0.50	−0.2735833	−0.2735857
15	0.75	−0.2284186	−0.2284204

7.

i	x_i	$\phi(x_i)$	$y(x_i)$
3	0.3	1.0408182	1.0408182
6	0.6	1.1065307	1.1065306
9	0.9	1.3065697	1.3065697

9. Un cambio de la variable $w = (x - a) / (b - a)$ origina el problema con valor en frontera

$$-\frac{d}{dw}(p((b-a)w+a)y') + (b-a)^2 q((b-a)w+a)y = (b-a)^2 f((b-a)w+a),$$

donde $0 < w < 1$, $y(0) = \alpha$ y $y(1) = \beta$. Entonces podemos utilizar el ejercicio 6.

13. Para $\mathbf{c} = (c_0, c_1, \ldots, c_{n+1})^t$ y $\phi(x) = \sum_{i=0}^{n+1} c_i\phi_i(x)$, tenemos

$$\mathbf{c}^t A\mathbf{c} = \int_0^1 p(x)\,[\phi'(x)]^2 + q(x)\,[\phi(x)]^2\,dx.$$

Pero $p(x) > 0$ y $q(x)[\phi(x)]^2 \geq 0$, por lo que $\mathbf{c}^tA\mathbf{c} \geq 0$, y puede ser 0, para $\mathbf{x} \neq 0$, sólo si $\phi'(x) \equiv 0$ en $[0, 1]$. Sin embargo, $\{\phi'_0, \phi'_1, \ldots, \phi'_{n+1}\}$ es linealmente independiente, por lo que $\phi'(x) \neq 0$ en $[0, 1]$ y $\mathbf{c}^tA\mathbf{c} = 0$ si y sólo si $\mathbf{c} = \mathbf{0}$.

Conjunto de ejercicios 12.1

1. El algoritmo de diferencias finitas para la ecuación de Poisson produce los siguientes resultados.

i	j	x_i	y_j	$w_{i,j}$	$u(x_i, y_j)$
1	1	0.5	0.5	0.0	0
1	2	0.5	1.0	0.25	0.25
1	3	0.5	1.5	1.0	1

3. El algoritmo de diferencias finitas para la ecuación de Poisson produce los siguientes resultados.

a. Se requieren 30 iteraciones:

i	j	x_i	y_j	$w_{i,j}$	$u(x_i, y_j)$
2	2	0.4	0.4	0.1599988	0.16
2	4	0.4	0.8	0.3199988	0.32
4	2	0.8	0.4	0.3199995	0.32
4	4	0.8	0.8	0.6399996	0.64

b. Se requieren 29 iteraciones:

i	j	x_i	y_j	$w_{i,j}$	$u(x_i, y_j)$
2	1	1.256637	0.3141593	0.2951855	0.2938926
2	3	1.256637	0.9424778	0.1830822	0.1816356
4	1	2.513274	0.3141593	−0.7721948	−0.7694209
4	3	2.513274	0.9424778	−0.4785169	−0.4755283

c. Se requieren 126 iteraciones:

i	j	x_i	y_j	$w_{i,j}$	$u(x_i, y_j)$
4	3	0.8	0.3	1.2714468	1.2712492
4	7	0.8	0.7	1.7509414	1.7506725
8	3	1.6	0.3	1.6167917	1.6160744
8	7	1.6	0.7	3.0659184	3.0648542

d. Se requieren 127 iteraciones:

i	j	x_i	y_j	$w_{i,j}$	$u(x_i, y_j)$
2	2	1.2	1.2	0.5251533	0.5250861
4	4	1.4	1.4	1.3190830	1.3189712
6	6	1.6	1.6	2.4065150	2.4064186
8	8	1.8	1.8	3.8088995	3.8088576

7. El potencial aproximado en algunos puntos típicos produce los siguientes resultados.

i	j	x_i	y_j	$w_{i,j}$
1	4	0.1	0.4	88
2	1	0.2	0.1	66
4	2	0.4	0.2	66

Conjunto de ejercicios 12.2

1. El algoritmo de las diferencias regresivas para la ecuación del calor produce los siguientes resultados.

a.

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
1	1	0.5	0.05	0.632952	0.652037
2	1	1.0	0.05	0.895129	0.883937
3	1	1.5	0.05	0.632952	0.625037
1	2	0.5	0.1	0.566574	0.552493
2	2	1.0	0.1	0.801256	0.781344
3	2	1.5	0.1	0.566574	0.552493

b.

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
1	1	1/3	0.05	1.59728	1.53102
2	1	2/3	0.05	−1.59728	−1.53102
1	2	1/3	0.1	1.47300	1.35333
2	2	2/3	0.1	−1.47300	−1.35333

3. El algoritmo de diferencias progresivas produce los siguientes resultados.

a. Para $h = 0.4$ y $k = 0.1$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	5	0.8	0.5	3.035630	0
3	5	1.2	0.5	−3.035630	0
4	5	1.6	0.5	1.876122	0

Para $h = 0.4$ y $k = 0.05$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	10	0.8	0.5	0	0
3	10	1.2	0.5	0	0
4	10	1.6	0.5	0	0

b. Para $h = \frac{\pi}{10}$ y $k = 0.05$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
3	10	0.94247780	0.5	0.4864823	0.4906936
6	10	1.88495559	0.5	0.5718943	0.5768449
9	10	2.82743339	0.5	0.1858197	0.1874283

c. Para $h = 0.2$ y $k = 0.04$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
4	10	0.8	0.4	1.166149	1.169362
8	10	1.6	0.4	1.252413	1.254556
12	10	2.4	0.4	0.4681813	0.4665473
16	10	3.2	0.4	−0.1027637	−0.1056622

d. Para $h = 0.1$ y $k = 0.04$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
3	10	0.3	0.4	0.5397009	0.5423003
6	10	0.6	0.4	0.6344565	0.6375122
9	10	0.9	0.4	0.2061474	0.2071403

5. El algoritmo de Crank-Nicolson produce los siguientes resultados.

a. Para $h = 0.4$ y $k = 0.1$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	5	0.8	0.5	8.2×10^{-7}	0
3	5	1.2	0.5	-8.2×10^{-7}	0
4	5	1.6	0.5	5.1×10^{-7}	0

Para $h = 0.4$ y $k = 0.05$

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	10	0.8	0.5	-2.6×10^{-6}	0
3	10	1.2	0.5	2.6×10^{-6}	0
4	10	1.6	0.5	-1.6×10^{-6}	0

b. Para $h = \frac{\pi}{10}$ y $k = 0.05$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
3	10	0.94247780	0.5	0.4926589	0.4906936
6	10	1.88495559	0.5	0.5791553	0.5768449
9	10	2.82743339	0.5	0.1881790	0.1874283

c. Para $h = 0.2$ y $k = 0.04$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
4	10	0.8	0.4	1.171532	1.169362
8	10	1.6	0.4	1.256005	1.254556
12	10	2.4	0.4	0.4654499	0.4665473
16	10	3.2	0.4	−0.1076139	−0.1056622

d. Para $h = 0.1$ y $k = 0.04$:

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
3	10	0.3	0.4	0.5440532	0.5423003
6	10	0.6	0.4	0.6395728	0.6375122
9	10	0.9	0.4	0.2078098	0.2071403

9. Para modificar el algoritmo 12.2, cambie lo siguiente:

Paso 7 Tome

$$t = jk;$$
$$z_1 = (w_1 + kF(h))/l_1.$$

Paso 8 Para $i = 2, \ldots, m-1$, tome

$$z_i = (w_i + kF(ih) + \lambda z_{i-1})/l_i$$

Para modificar el algoritmo 12.3, cambie lo siguiente:

Paso 7 Tome

$$t = jk;$$
$$z_i = \left[(1-\lambda)w_1 + \frac{\lambda}{2}w_2 + kF(h)\right]\Big/l_1.$$

Paso 8 Para $i = 2, \ldots, m-1$, tome

$$z_i = \left[(1-\lambda)w_i + \frac{\lambda}{2}(w_{i+1} + w_{i-1} + z_{i-1}) + kF(ih)\right]\Big/l_i.$$

13. a. La temperatura aproximada en algunos puntos típicos se incluyen en la tabla anexa.

i	j	r_i	t_j	$w_{i,j}$
1	20	0.6	10	137.6753
2	20	0.7	10	245.9678
3	20	0.8	10	340.2862
4	20	0.9	10	424.1537

b. La deformación es aproximadamente $I = 1242.537$.

Conjunto de ejercicios 12.3

1. El algoritmo de diferencias finitas para la ecuación de onda produce los siguientes resultados.

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	4	0.25	1.0	−0.7071068	−0.7071068
3	4	0.50	1.0	−1.0000000	−1.0000000
4	4	0.75	1.0	−0.7071068	−0.7071068

3. El algoritmo de diferencias finitas para la ecuación de onda con $h = \frac{\pi}{10}$ y $k = 0.05$ produce los siguientes resultados.

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	10	$\frac{\pi}{5}$	0.5	0.5163933	0.5158301
5	10	$\frac{\pi}{2}$	0.5	0.8785407	0.8775826
8	10	$\frac{4\pi}{5}$	0.5	0.5163933	0.5158301

El algoritmo de diferencias finitas para la ecuación de onda con $h = \frac{\pi}{20}$ y $k = 0.1$ produce los siguientes resultados.

i	j	x_i	t_j	w_{ij}
4	5	$\frac{\pi}{5}$	0.5	0.5159163
10	5	$\frac{\pi}{2}$	0.5	0.8777292
16	5	$\frac{4\pi}{5}$	0.5	0.5159163

El algoritmo de diferencias finitas para la ecuación de onda con $h = \frac{\pi}{20}$ y $k = 0.05$ produce los siguientes resultados.

i	j	x_i	t_j	w_{ij}
4	10	$\frac{\pi}{5}$	0.5	0.5159602
10	10	$\frac{\pi}{2}$	0.5	0.8778039
16	10	$\frac{4\pi}{5}$	0.5	0.5159602

5. El algoritmo de diferencias finitas para la ecuación de onda produce los siguientes resultados.

i	j	x_i	t_j	w_{ij}	$u(x_i, t_j)$
2	3	0.2	0.3	0.6729902	0.61061587
5	3	0.5	0.3	0	0
8	3	0.8	0.3	−0.6729902	−0.61061587

7. a. La presión de aire en el tubo abierto es $p(0.5, 0.5) \approx 0.9$ y $p(0.5, 1.0) \approx 2.7$.

b. La presión de aire en el tubo cerrado es $p(0.5, 0.5) \approx 0.9$ y $p(0.5, 1.0) \approx 0.9187927$.

Conjunto de ejercicios 12.4

1. Con $E_1 = (0.25, 0.75)$, $E_2 = (0, 1)$, $E_3 = (0.5, 0.5)$ y $E_4 = (0, 0.5)$, las funciones base son

$$\phi_1(x, y) = \begin{cases} 4x & \text{en } T_1 \\ -2 + 4y & \text{en } T_2, \end{cases}$$

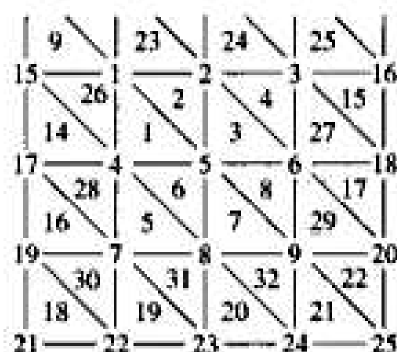
$$\phi_2(x, y) = \begin{cases} -1 - 2x + 2y & \text{en } T_1 \\ 0 & \text{en } T_2, \end{cases}$$

$$\phi_3(x, y) = \begin{cases} 0 & \text{en } T_1 \\ 1 + 2x - 2y & \text{en } T_2, \end{cases}$$

$$\phi_4(x, y) = \begin{cases} 2 - 2x - 2y & \text{en } T_1 \\ 2 - 2x - 2y & \text{en } T_2, \end{cases}$$

y $\gamma_1 = 0.323825$, $\gamma_2 = 0$, $\gamma_3 = 1.0000$ y $\gamma_4 = 0$.

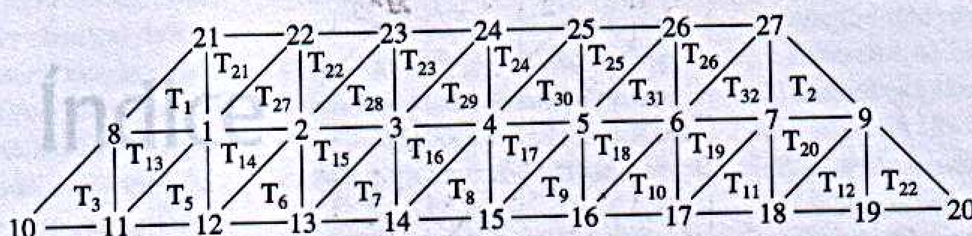
3. El algoritmo de elementos finitos con $K = 8$, $N = 8$, $M = 32$, $n = 9$, $m = 25$ y $NL = 0$ produce los siguientes resultados. (Véase el diagrama.)



$$\gamma_1 = 0.511023$$
$$\gamma_2 = 0.720476$$
$$\gamma_3 = 0.507899$$
$$\gamma_4 = 0.720476$$
$$\gamma_5 = 1.01885$$
$$\gamma_6 = 0.720476$$
$$\gamma_7 = 0.507896$$
$$\gamma_8 = 0.720476$$
$$\gamma_9 = 0.511023$$
$$\gamma_i = 0 \quad 10 \le i \le 25$$

$$u(0.125, 0.125) \approx 0.614187$$
$$u(0.125, 0.25) \approx 0.690343$$
$$u(0.25, 0.125) \approx 0.690343$$
$$u(0.25, 0.25) \approx 0.720476$$

5. El algoritmo de elementos finitos con $K = 0$, $N = 12$, $M = 32$, $n = 20$, $m = 27$ y $NL = 14$ produce los siguientes resultados. (Véase el diagrama.)



$\gamma_1 = 21.40335$	$\gamma_8 = 24.19855$	$\gamma_{15} = 20.23334$	$\gamma_{22} = 15$
$\gamma_2 = 19.87372$	$\gamma_9 = 24.16799$	$\gamma_{16} = 20.50056$	$\gamma_{23} = 15$
$\gamma_3 = 19.10019$	$\gamma_{10} = 27.55237$	$\gamma_{17} = 21.35070$	$\gamma_{24} = 15$
$\gamma_4 = 18.85895$	$\gamma_{11} = 25.11508$	$\gamma_{18} = 22.84663$	$\gamma_{25} = 15$
$\gamma_5 = 19.08533$	$\gamma_{12} = 22.92824$	$\gamma_{19} = 24.98178$	$\gamma_{26} = 15$
$\gamma_6 = 19.84115$	$\gamma_{13} = 21.39741$	$\gamma_{20} = 27.41907$	$\gamma_{27} = 15$
$\gamma_7 = 21.34694$	$\gamma_{14} = 20.52179$	$\gamma_{21} = 15$	

$$u(1, 0) \approx 22.92824$$
$$u(4, 0) \approx 22.84663$$
$$u\left(\frac{5}{2}, \frac{\sqrt{3}}{2}\right) \approx 18.85895$$

Índice

La séptima edición de Análisis numérico incorpora los métodos de aproximación numérica más modernos, por ejemplo, el de gradiente conjugado y el de homotopía; este último permite que el estudiante tome conciencia de la importancia en la continuidad del desarrollo en esta área. Los más de 2,000 ejercicios han sido probados en clase e incluyen desde las aplicaciones más simples de métodos y algoritmos, hasta generalizaciones de la teoría en varias disciplinas.

THOMSON LEARNING

MÉXICO Y AMÉRICA CENTRAL
Tel. 52(55) 5281-2906
Fax 52(55) 5280-8970
editor@thomsonlearning.com.mx
México, D.F., MÉXICO

AMÉRICA DEL SUR
Tels. (5411) 4833-3838/3883
Fax (5411) 4831-0764
thomson@thomsonlearning.com.ar
Buenos Aires, ARGENTINA

EL CARIBE
Tel. (787) 758-7580
Fax (787) 758-7573
thomson@coqui.net
Hato Rey, PUERTO RICO

PACTO ANDINO
Tel. (571) 340-9470
Fax (571) 340-9475
clithomson@andinet.com
Bogotá, COLOMBIA

ESPAÑA
Tel. (3491) 446-3350
Fax (3491) 445-6218
clientes@paraninfo.es
Madrid, ESPAÑA

